

Voiceless Speech Recognition Using Dynamic Visual Speech Features

Wai Chee Yau

Dinesh Kant Kumar

Sridhar Poosapadi Arjunan

School of Electrical and Computer Engineering, RMIT University
GPO Box 2476V Melbourne, Victoria 3001, Australia
Email: waichee@ieee.org

Abstract

This paper describes a voiceless speech recognition technique that utilizes dynamic visual features to represent the facial movements during phonation. The dynamic features extracted from the mouth video are used to classify utterances without using the acoustic data. The audio signals of consonants are more confusing than vowels and the facial movements involved in pronunciation of consonants are more discernible. Thus, this paper focuses on identifying consonants using visual information. This paper adopts a visual speech model that categorizes utterances into sequences of smallest visually distinguishable units known as visemes. The viseme model used is based on the viseme model of Moving Picture Experts Group 4 (MPEG-4) standard. The facial movements are segmented from the video data using motion history images (MHI). MHI is a spatio-temporal template (grayscale image) generated from the video data using accumulative image subtraction technique. The proposed approach combines discrete stationary wavelet transform (SWT) and Zernike moments to extract rotation invariant features from the MHI. A feedforward multilayer perceptron (MLP) neural network is used to classify the features based on the patterns of visible facial movements. The preliminary experimental results indicate that the proposed technique is suitable for recognition of English consonants.

Keywords: visual speech recognition, wavelet transform, feedforward neural network.

1 Introduction

Speech recognition has been an important research subject that spans across multiple disciplines such as human-computer interaction (HCI), signal processing, linguistic and machine learning. Enormous research efforts are put into developing intelligent machines that are capable of comprehending utterances. Such speech-based devices are useful as they provide the flexibility for users to control computers using human speech.

However, the performance of the current speech recognition systems are still far behind as compare to human's cognitive ability in perceiving and understanding speech (Lippmann 1997). The main difficulty of the conventional speech recognition techniques based on audio signals is that such systems are sensitive to signal strength, ambient noise

and acoustic conditions. To overcome this limitation, the non acoustic speech modalities can be used to complement the audio signals. There are a number of options available such as visual (Goecke & Millar 2003, Potamianos, Neti, Huang, Connell, Chu, Libal, Marcheret, Haas & Jiang 2004), recording of vocal cords movements through electroglottograph (EGG) (Dikshit & R.W.Schubert 1995), mechanical sensing of facial movement and movement of palate, recording of facial muscle activity (Arjunan, Kumar, Yau & Weghorn 2006), facial plethysmogram and measuring the intra-oral pressure (Soquet, Saerens & Lecuit 1999). Vision-based speech recognition techniques are least intrusive and non invasive and this paper reports on such a technique for HCI application.

In our normal communication, the visual modality of speech is often incorporated into audio speech recognition (ASR) systems because the visual speech signals are invariant to acoustic noise and style of speech. Such systems that combine the audio and visual modalities to identify utterances are known as audio-visual speech recognition (AVSR) systems. AVSR systems can enhance the performance of the conventional ASR system especially under noisy condition (Chen 2001). Research where these AVSR systems are being made more robust, and able to recognize complex speech patterns of multiple speakers are being reported (Potamianos et al. 2004, Liang, Liu, Zhao, Pi & Nefian 2002). While AVSR systems are useful for applications such as for telephony in noisy environment, these are not suitable for people with speech impairment that have difficulty in producing speech sounds. AVSR systems are also not useful in situations where it is essential to maintain silence. Thus, the need for a voiceless, visual-only communication system arises. Such a system is also commonly known as lipreading or visual speech recognition or speechreading system.

Speechreading systems use the visual information extracted from the image sequence of the mouth to identify utterances. The visual speech information refers to the movement of the speech articulators such as the lips, facial muscles, tongue, teeth and jaw of the speaker. The complex range of reproducible sounds produced by people is a clear demonstration of the dexterity of the human mouth and lips- the key speech articulators. The possible advantages of such voiceless systems are (i) not sensitive to audio noise and change in acoustic conditions (ii) does not require the user to make a sound and (iii) suitable for users with speech impairment.

The visual cues contain far less classification power for speech compared to audio data and hence it is to be expected that speechreading systems would have only a small vocabulary. Such systems are also known to be user dependent, and hence it is important for such a system to be easy to train for a new user. And

because there is no audio cue, it is highly desirable that the system provide the user with active feedback to avoid any error in communication.

The main limitation with current speechreading systems is that these systems adopt a 'one size fits all' approach. Due to the large variation in the way people speak English, especially if we transgress the national and cultural boundaries, these have very high error rate, with error of the order of 90% for large vocabulary systems (Potamianos, Neti, Gravier & Senior 2003, Hazen 2006) and error rates in the range of 55% to 90% for small vocabulary system (Matthews, Cootes, Cox, Harvey & Bangham 1998), which demonstrates the inability of these systems to be used as voiceless, speech-controlled human computer interfaces.

What is required is a speaker-dependent system that is easy to train for individual users, with low computational complexity and can provide active feedback to the user. The system needs to be robust under changing conditions such as angle and distance of the camera, and insensitive to factors such as different skin color, texture and rapidity of speech of the speaker.

To achieve the above mentioned goals, this paper proposes a system where the camera is attached in place of the microphone to the commonly available head-sets. The advantage of this is that using this, it is no longer required to identify the region of interest, reducing the computation required. The video processing proposed is the use of accumulative image subtraction technique to directly segment the facial movements of the speaker. The proposed technique uses dynamic visual speech features based on the movements of the lower face region such as movements of the mouth, jaw and facial muscles. The proposed motion segmentation approach is based on the use of motion history images (MHI) where the video data is multiply with a ramp function and temporally integrated with greater weight to the recent movements. The resultant MHI is a 2-D grayscale image which is suitable for representing short duration complex movements of the lower face. Section 2 discusses on related work in the field of speechreading and Section 3 describes our proposed approach. Section 4 presents the methodology and Section 5 reports on the results of the initial experiments. Section 6 presents the discussion and findings based on the experimental results and Section 7 discusses the recommendations for possible future work.

2 Related Work

Numerous speechreading techniques are reported in the literature and comprehensive reviews on speech recognition research can be found in (Potamianos et al. 2003, Chen 2001, Stork & Hennecke 1996). Visual features used in speechreading systems can be divided into two main categories - shape-based and intensity-based. The shape-based features rely on the geometric shape of the mouth and lips and can be represented by a small number of parameters. The first speechreading system was proposed by Petajan (Petajan 1984) using shape-based features such as height, width and area of the mouth derived from the binary mouth images. Shape based features based on 3D coordinates of feature points (lip corners and midpoints of upper and lower lip) are extracted from stereo images in (Goecke & Millar 2003). Lip contours extracted using deformable template techniques such as active shape models (ASM) are used as visual speech features in (Matthews et al. 1998, Perez, Frangi, Solano & Lukas 2005). ASM obtains the lip information by fitting a statistical shape model of the

lip to the video frames. While such top-down, model-based approaches are less sensitive to the view angle of the camera and image noise, they rely only on the shape of the lip contours and do not contain information of other speech articulators. An extension to the ASM technique is active appearance model (AAM) that combines the shape model with a statistical model of the grey levels in the mouth region. The performance of AAM is demonstrated to outperform ASM in lip tracking (Matthews et al. 1998). However, both AAM and ASM techniques are sensitive to tracking error and modelling error.

Intensity-based features are derived directly from the pixel intensity values of the image around the mouth area (Liang et al. 2002, Potamianos et al. 2004, Hazen 2006, Saenko, Darrell & Glass 2004). Such features are extracted using bottom-up approach. The advantage of intensity-based systems is that accurate tracking and modelling of the lips are not required as opposed to model-based systems. The training of the statistical model of the lips is also not necessary for intensity-based approach thereby reducing the computational complexity of the systems. Intensity-based features are capable of representing visual information within the mouth cavity and also surrounding face region that are not represented in the high-level, shape-based features and lip contours (Potamianos et al. 2004). Nonetheless, the intensity-based features have much higher dimensionality if taking directly all the pixels from the mouth images. Dimensionality reduction or feature extraction techniques such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA) can be applied on the images to reduce the dimension of such features. The intensity-based features are demonstrated to yield better performance than shape-based features extracted using ASM and AAM algorithms in (Matthews et al. 1998). Similarly, intensity-based features using Discrete Cosine Transform (DCT) is also shown to outperform model-based features obtained using ASM algorithm in (Perez et al. 2005). This paper reports on the use of intensity-based features extracted from the MHI to represent facial movements for consonants recognition.

3 Theory

3.1 Visual Speech Model - Viseme Model

Human speech is organized as sequences of basic unit of speech sounds known as phoneme. Phonemes can be further dichotomized into vowels and consonants. The audio signals of consonants are less distinguishable than vowels (Chen 2001). Hence, the visual speech information is crucial in differentiating the consonants, especially in conditions where the acoustic signal strength is low or contaminated by noise. This paper focuses on the recognition of consonants due to the fact that consonants are easier to "see" and harder to "hear" than vowels (Kaplan, Bally & Garretson 1999). The pronunciation of vowels are produced with an open vocal tract whereas the production of consonants involve constrictions at certain part of the vocal tract by the speech articulators. Hence, the facial movements involved in pronunciation of consonants are more discernible than vowels. To represent the different facial movements when uttering consonants, a visual speech model is required.

This paper uses visemes to model visual speech. The motivation of using viseme as the recognition unit is because visemes can be concatenated to form words and sentences, thus providing the flexibility for the proposed visual speech recognition system to be extended into a large vocabulary system. The to-

tal number of visemes is much less than phonemes because speech is only partially visible (Hazen 2006). While the video of the speaker’s face shows the movement of the lips and jaw, the movements of other articulators such as tongue and vocal cords are often not visible. Hence, each viseme can correspond to more than one phoneme, resulting in a many-to-one mapping of phonemes-to-visemes.

Various viseme models had been proposed for AVSR applications (Hazen, Saenko, La & Glass 2004, Potamianos et al. 2004, Gordan, Kotropoulos & Pitas 2002). There is no definite consensus about how the sets of visemes in English is constituted (Hazen 2006). The number of visemes for English varies depending on factors such as the geographical location, culture, education background and age of the speaker. The geographic differences in English is most obvious where the sets of phonemes and visemes changes for different countries and even for areas within the same country. It is difficult to determine an optimal and universal viseme set that is suitable for all users. This paper adopts the viseme model established for facial animation applications by an international audiovisual object-based video representation standard known as MPEG-4. The motivation of using this model is because this enable the proposed system to be coupled with any MPEG-4 supported facial animation systems to form an interactive speech recognition and synthesis human computer interface. Based on the MPEG-4 viseme model, there is nine visemes associated with all English consonants. This paper adopts this nine visemes to represent the different facial movements when pronouncing consonants. The consonants chosen for experiments for each of the nine visemes are highlighted in bold fonts in Table 1.

Viseme Number	Phonemes	Example words
1	/p/,/b/,/m/	put, bed, me
2	/f/,/v/	far, voice
3	/th/,/dh/	think, that
4	/t/,/d/	tick, door
5	/k/,/g/	kick, gate
6	/sh/, /j/, /ch/	she, join, chair
7	/s/,/z/	sick, zeal
8	/n/,/l/	new, less
9	/r/	rest

Table 1: Viseme model of the MPEG-4 standard for English consonants.

3.2 Segmentation of the Facial Movements

In the proposed approach, the dynamic visual speech features which comprise of the facial movements of the speaker are segmented from the video data using a view-based approach named motion history images (MHI) (Bobick & Davis 2001). MHI is a spatial-temporal template that shows where and when movement of speech articulators (lips, teeth, jaw, facial muscles and tongue) occurs in the image sequence. MHI is generated using difference of frames (DOF) from the video of the speaker. Accumulative image subtraction is applied on the image sequence by subtracting the intensity values between successive frames to generate the difference of frames (DOFs). The delimiters for the start and stop of the motion are manually inserted into the image sequence of every articulation. The MHI of the video of the lips would have pixels corresponding to the more recent mouth movement brighter with larger intensity values. The intensity value of the MHI at pixel location

(x, y) of time t (or the t^{th} frame) is defined by

$$MHI_t = \max \bigcup_{t=1}^{N-1} B(x, y, t) \times t \quad (1)$$

N is the total number of frames used to capture the mouth motion. $B(x, y, t)$ is the binarisation of the DOF using the threshold a and $B(x, y, t)$ is given by

$$B(x, y, t) = \begin{cases} 1 & \text{if } Diff(x, y, t) \geq a, \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

a is the predetermined threshold for binarisation of the DOF represented as $Diff(x, y, t)$. The value for the fixed threshold, a is optimized through experimentation. The DOF of the t^{th} frame is defined as

$$Diff(x, y, t) = |I(x, y, t) - I(x, y, t - 1)| \quad (3)$$

$I(x, y, t)$ represents the intensity value of pixel location with coordinate (x, y) at the t^{th} frame of the image sequence. In Eq. (1), the binarised version of the DOF is multiplied with a linear ramp of time to implicitly encode the timing information of the motion into the MHI (Kumar & Kumar 2005). By computing the MHI values for all the pixels coordinates (x, y) of the image sequence using Eq. (1) will produce a scalar-valued grayscale image (MHI) where the brightness of the pixels indicates the recency of motion in the image sequence. The proposed motion segmentation approach is computationally simple and is suitable for real time implementation. Figure 1 shows examples of 3 MHIs generated from the video of the speaker and Figure 2 illustrates the 9 MHIs that form the viseme model of MPEG-4 for English consonants used in the experiments.

The motivation of using MHI in visual speech recognition is the ability of MHI to remove static elements from the sequence of images and preserve the short duration facial movements. MHI is also invariant to the skin color of the speakers due to the DOF and image subtraction process involved in the generation of MHI.

3.2.1 Variation in Speed of Speech

The speed of phonation of the speaker might varies for each pronunciation of a phone. Hence, the speed of the mouth movements when the speaker is pronouncing a consonant might be different for each video recording. The variation in the speed of utterance results in the variation of the overall duration and there maybe variation in the microphases of the utterances. The details of such variations are difficult to model due to the large inter-subject and inter-experiment variations. This paper suggests a model to approximate such variations by normalizing the overall duration of the utterance. This is achieved by normalizing the intensity values of the MHI to in between 0 and 1 to minimize the difference in MHIs produced from video data of different rapidity of speech.

3.2.2 Issues Related to the Segmentation of the Facial Movements

MHI is a view sensitive motion representation technique. Therefore the MHI generated from the sequence of images is dependent on factors such as:

1. position of the speaker’s mouth normal to the camera optical axis
2. orientation of the speaker’s face with respect to the video camera







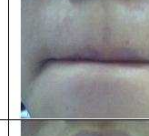
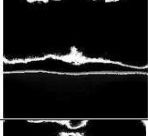




Consonants	Start Frame	Middle Frame	End Frame	MHI
/v/				
/m/				
/g/				

Figure 1: Examples of MHI generated from video of a speaker uttering three different consonants.

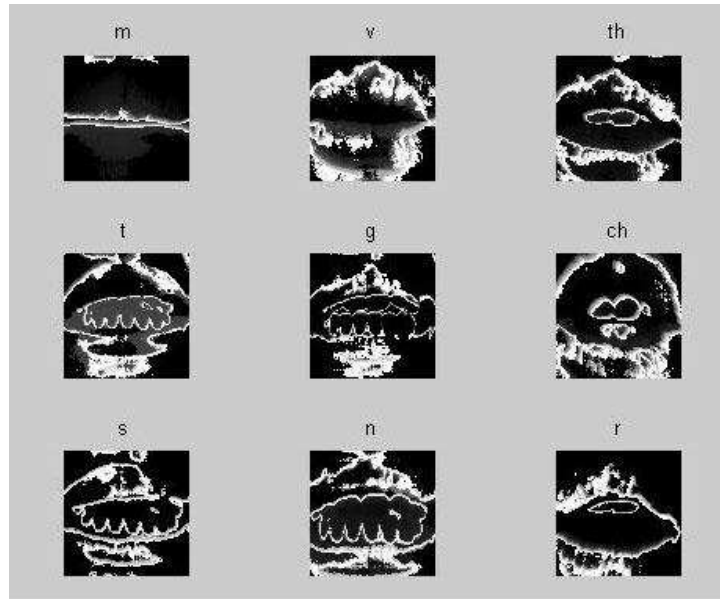


Figure 2: MHI of the 9 consonants from the MPEG-4 viseme model

3. distance of the speaker's mouth from the camera (which changes the scale/size of the mouth in the video data)
4. small variation of the mouth movement of the speaker while uttering the same consonant

This paper proposes the use of approximate image of discrete stationary wavelet transform (SWT) to obtain a time-frequency representation of the MHI that is insensitive to small variations of the mouth and lip movement. The proposed technique adopts Zernike moments as the region-based features to represent the SWT approximate image of the MHI to further reduce the dimension of the data. Zernike moments are chosen because they can be normalized to achieve rotation invariance.

3.2.3 Discrete Stationary Wavelet Transform

This paper proposes the use of discrete stationary wavelet transform (SWT) to obtain a transform representation of the MHI that is insensitive to small variations of the mouth and lip movement. While the classical discrete wavelet transform (DWT) is suitable for this, DWT results in translation variance (Mallat

1998) where a small shift of the image in the space domain will yield very different wavelet coefficients. The translation sensitivity of DWT is caused by the aliasing effect that occurs due to the downsampling of the image along rows and columns (Simoncelli, Freeman, Adelson & Heeger 1992). SWT restores the translation invariance of the signal by omitting the downsampling process of DWT, and results in redundancies.

2-D SWT at level 1 is applied on the MHI to produce a spatial-frequency representation of the MHI. The 2-D SWT is implemented by applying 1-D SWT along the rows of the image followed by 1-D SWT along the columns of the image. SWT decomposition of the MHI generates four images, namely approximation (LL), horizontal detail coefficients (LH), vertical detail coefficients (HL) and diagonal detail coefficients (HH) through iterative filtering using low pass and high pass filters. The approximate image is the smoothed version of the MHI and carries the highest amount of information content among the four images. LH, HL and HH sub images show the fluctuations of the pixel intensity values in the horizontal, vertical and diagonal directions respectively. The image moments features are computed from the ap-

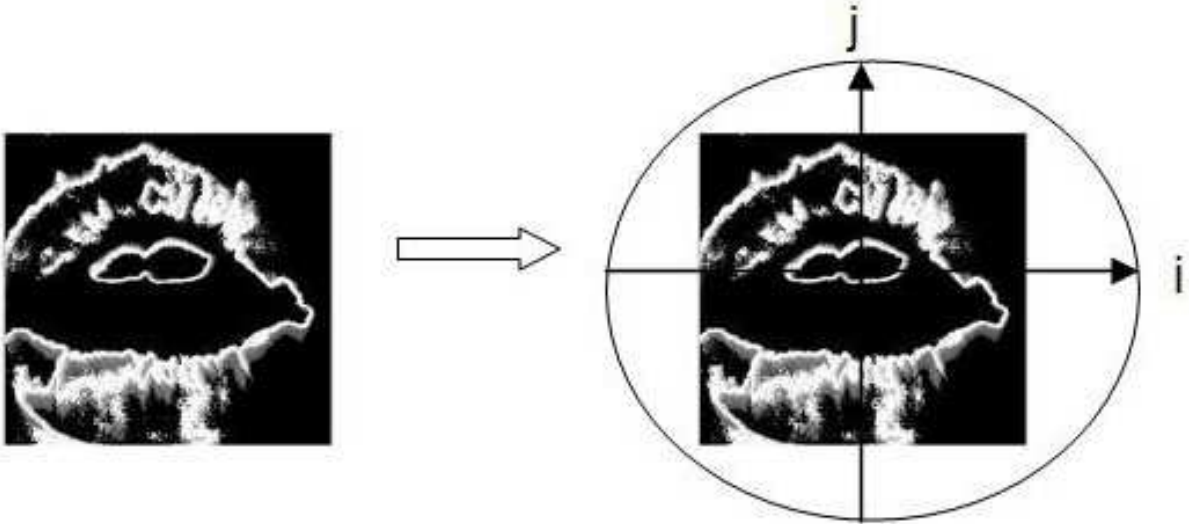


Figure 3: The square-to-circular transformation of the SWT approximation of MHI

proximate sub image.

Haar wavelet has been selected due to its spatial compactness and localization property. Another advantage is the low mathematical complexity of this wavelet. Compact features have to be extracted from the approximation (LL) to further reduce the size of the data. Since the gray levels of MHI are the temporal descriptors of motion occurring in the image sequence, thus it is intuitive to use global region-based feature descriptors to represent the approximation of the MHI. The proposed technique adopts Zernike moments as the region-based features to represent the SWT approximate image of the MHI.

3.3 Visual Speech Features - Zernike Moments

Zernike moments are image moments commonly used in recognition of image patterns (Khontazad & Hong 1990, Teague 1980). Zernike moments have been demonstrated to outperformed other image moments such as geometric moments, Legendre moments and complex moments in terms of sensitivity to image noise, information redundancy and capability for image representation (Teh & Chin 1988). The proposed technique uses Zernike moments as visual speech features to represent the SWT approximate image of the MHI.

Zernike moments are computed by projecting the image function $f(x, y)$ onto the orthogonal Zernike polynomial V_{nl} of order n with repetition l is defined within a unit circle (i.e.: $x^2 + y^2 \leq 1$) as follows:

$$V_{nl}(\rho, \theta) = R_{nl}(\rho)e^{-j\hat{l}\theta}; \hat{j} = \sqrt{-1} \quad (4)$$

where R_{nl} is the real-valued radial polynomial

The main advantage of Zernike moments is the simple rotational property of the features (Khontazad & Hong 1990). Zernike moments are also independent features due to the orthogonality of the Zernike polynomial V_{nl} (Teh & Chin 1988). $|l| \leq n$ and $(n - |l|)$ is even. Zernike moments Z_{nl} of order n and repetition l is given by

$$Z_{nl} = \left[\frac{n+1}{\pi} \right] \int_0^{2\pi} \int_0^{\infty} [V_{nl}(\rho, \theta)] f^*(\rho, \theta) \rho d\rho d\theta \quad (5)$$

$f(\rho, \theta)$ is the intensity distribution of the approximate image of MHI mapped to a unit circle of radius ρ and angle θ where $x = \rho \cos\theta$ and $y = \rho \sin\theta$.

For the Zernike moments to be orthogonal, the approximate image of the MHI is scaled to be within a unit circle centered at the origin. The unit circle is bounded by the square approximate image of the MHI. The center of the image is taken as the origin and the pixel coordinates are mapped to the range of the unit circle i.e.: $x^2 + y^2 \leq 1$. Figure 3 shows the square-to-circular transformation performed for the computation of the Zernike moments that transform the square image function ($f(x, y)$) in terms of the x-y axes to a circular image function ($f(\rho, \theta)$) in terms of the i-j axes.

To illustrate the rotational characteristics of Zernike moments, consider β as the angle of rotation of the image. The resulting rotated Zernike moment Z'_{nl} is

$$Z'_{nl} = Z_{nl}e^{-il\beta} \quad (6)$$

Z_{nl} is the Zernike moment of the original image. Eq. (6) demonstrates that rotation of an image results in a phase shift on the Zernike moments (Teague 1980). The absolute value of Zernike moments are rotation invariant (Khontazad & Hong 1990) as shown in the equation below

$$|Z'_{nl}| = |Z_{nl}| \quad (7)$$

This paper uses the absolute value of the Zernike moments, $|Z'_{nl}|$ as the rotation invariant features of the SWT of MHI. By including higher order moments, more information of the MHI can be represented by the Zernike moments features. However, this inherently increases the size of the features and makes it prone to noise. An optimum number of Zernike moments need to be selected to trade-off between the dimensionality of the feature vectors and the amount of information represented by the features. 49 Zernike moments that comprise of 0th order moments up to 12th order moments have been used as features to represent the approximate image of the MHI for each consonant. Table 2 lists the 49 Zernike moments used in the experiments.

Order	Moments	No. of Moments
0	Z_{00}	1
1	Z_{11}	1
2	$Z_{20} Z_{22}$	2
3	$Z_{31} Z_{33}$	2
4	$Z_{40} Z_{42} Z_{44}$	3
5	$Z_{51} Z_{53} Z_{55}$	3
6	$Z_{60} Z_{62} Z_{64} Z_{66}$	4
7	$Z_{71} Z_{73} Z_{75} Z_{77}$	4
8	$Z_{80} Z_{82} Z_{84} Z_{86} Z_{88}$	5
9	$Z_{91} Z_{93} Z_{95} Z_{97} Z_{99}$	5
10	$Z_{10,0} Z_{10,2} Z_{10,4} Z_{10,6} Z_{10,8} Z_{10,10}$	6
11	$Z_{11,1} Z_{11,3} Z_{11,5} Z_{11,7} Z_{11,9} Z_{11,11}$	6
12	$Z_{12,0} Z_{12,2} Z_{12,4} Z_{12,6} Z_{12,8} Z_{12,10} Z_{12,12}$	7

Table 2: List of the 49 Zernike Moments and Their Corresponding Number of Features From Order Zero to Order Twelve

3.4 Classification Using Feedforward Neural Network

There are a number of possible classifiers that maybe suitable for such a system. The selection of the appropriate classifier would require statistical analysis of the data that would also identify the features that are irrelevant. Supervised neural network approach lends itself for identifying the separability of data even when the statistical properties and the types of separability (linear or nonlinear) is not known and without even requiring the estimating of the kernel. While it may be suboptimum, it is an easy tool to implement as a first step.

This paper presents the use of artificial neural network (ANN) to classify Zernike moments features into one of the class of consonants. ANN has been selected because it can solve complicated problems where the description for the data is not easy to compute. The other advantage of the use of ANN is its fault tolerance and high computation rate due to the massive parallelism of its structure(Kulkarni 1994). The functionality of the ANN to be less dependent on the underlying distribution of the classes as opposed to other classifiers such as Bayesian classifier and Hidden Markov Models(HMM) is yet another advantage for using ANN in this application(Stork & Hennecke 1996).

A supervised feed-forward multilayer perceptron (MLP) ANN classifier with back propagation(BP) learning algorithm is integrated in the visual speech recognition system described in this paper. The ANN is provided with a number of training vectors for each class during the training phase. MLP ANN was selected due to its ability to work with complex data compared with a single layer network. Due to the multilayer construction, such a network can be used to approximate any continuous functional mapping(Bishop 1995). This paper proposes the use of a three-layer network with BP learning algorithm to classify the visual speech features. The advantage of using BP learning algorithm is that the inputs are augmented with hidden context units to give feedback to the hidden layer and extract features of the data from the training events(Haung 2001). Trained ANNs have very fast classification speed(Freeman & Skapura 1991) thus making them an appropriate classifier choice for real time visual speech recognition applications. Figure 4 shows the overall block diagram of the proposed technique.

4 Experiments

Experiments were conducted to evaluate the performance of the proposed visual speech recognition techniques in classifying English consonants. The experiments were approved by the Human Experiments Ethics Committee of the university. Nine consonants that form the viseme model of English consonants according to the MPEG-4 standard are tested in the experiment. The nine consonants tested (/m/, /v/, /th/, /t/, /g/, /ch/, /s/, /n/ and /r/) were highlighted in bold in Table 1.

4.1 Video Recording and Processing

Video data was recorded from one speaker using an inexpensive web camera in a typical office environment. This was done towards having an inexpensive and practical voiceless communication system using low resolution video recordings. The video camera focused on the mouth region of the speaker and the camera was kept stationary throughout the experiment. The following factors were kept the same during the recording of the videos : window size and view angle of the camera, background and illumination. 20 video data of size 240 x 240 was recorded for each of the nine consonants. Thus, a total of 180 video data was created. The video data was stored as true color (.AVI) files and every AVI file had a duration of two seconds to ensure that the speaker had sufficient time to utter each of the consonant. The frame rate of the AVI files was 30 frames per second. One MHI was generated from each of the AVI file. An example of MHI for each of the nine consonants are shown in Figure 3.

4.2 Features Extraction

SWT at level-1 using Haar wavelet was applied on the MHIs and the approximate image (LL) was used for analysis. Zernike moments are computed from circular region of interest while the MHI is a square image. Hence, square-to-circular transformation of the SWT approximate image of the MHI had been done to compute the orthogonal Zernike moments features. 49 Zernike moments that comprise of 0th order moments up to 12th order moments have been used as features to represent the SWT approximate image of the MHI for each consonants.

4.3 Classification

The next step of the experiments was to classify the features using artificial neural network(ANN), which

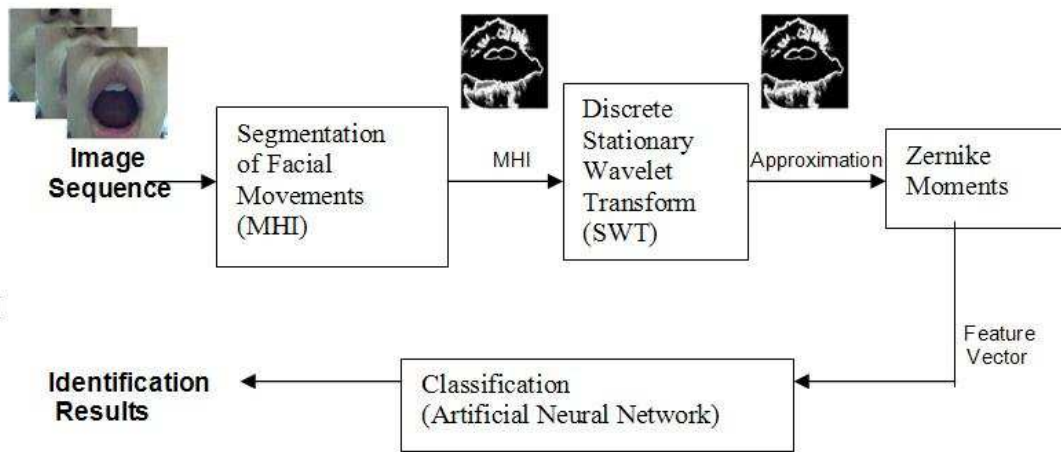


Figure 4: Block diagram of the proposed visual speech recognition approach.

Viseme	Recognition Rate
/m/	100%
/v/	87%
/th/	65%
/t/	74%
/g/	85%
/ch/	91%
/s/	93%
/n/	74%
/r/	93%

Table 3: Mean Classification Accuracies for 9 Visemes(Consonants).

can learn patterns of features with nonlinear separation. The Zernike moments features were fed to ANN to classify the features into one of the consonants. Multilayer perceptron (MLP) ANN with back-propagation (BP) learning algorithm was employed in the proposed system. The architecture of the ANN consisted of two hidden layers. The size of the input layer of the ANN was chosen to be same as the size of the features which was 49 nodes. The size of the output layer of the ANN was 9 which corresponded to the number of visemes (classes) available. The total numbers of hidden nodes was 140 which was determined iteratively through experimentation. Sigmoid function was the threshold function and the type of training algorithm for the ANN was gradient descent and adaptive learning with momentum with a learning rate of 0.05 to reduce chances of local minima. In the experiments, Zernike moments features of 10 MHIs of each consonants were used to train the ANN. The remaining 10 MHIs (that were not used in training the ANN) were presented to the ANN to test the ability of the trained ANN in recognizing the nine consonants. The statistical mean and variance of the classification accuracies of the data are determined by repeating the experiments 10 times. For each repetition of the experiment, the 10 test samples for each consonants were selected randomly with different combinations (permutations) to train the ANN and the remaining 10 MHIs were used as test samples.

5 Results and Observations

The experiments have tested the robustness of the use of MHI features to identify the human speech visemes with a feedforward neural network as the classifier. The ANN used was trained for an individual subject. The mean recognition accuracies of the ANN for the 10 repetitions of the experiments are tabulated in Table 3. From this table, it is observed that the mean success rate for identifying the viseme based consonants is 84.7% with a standard deviation of 2.8%.

6 Discussion

The results indicate that the proposed technique based on dynamic visual speech information (facial movements) is suitable for consonants recognition. The results indicate that the different patterns of facial movements can be used to classify the 9 visemes of English consonants based on the MPEG-4 viseme model.

The good results demonstrate the ability of ANN to learn the patterns of the facial movement features. From the results, it is observed that a small number of samples are sufficient to suitably train the ANN based system, indicating the sufficient compactness of each class of the data.

One of the possible reason for the misclassifications of the test samples by the ANN can be attributed to the inability of vision-based technique to capture the occluded articulators movements. Example, the movement of the tongue within the mouth cavity is not visible (occluded by the teeth) in the video data during the pronunciation of /n/. Thus, the resultant MHI of /n/ does not contain information on the tongue movement.

While the error rates of the experiments are much lower than the 90% error reported by (Potamianos et al. 2003, Hazen 2006), the authors would like to point out that it is not appropriate to compare our results with other related work as this system has only been tested using a small vocabulary consisting of discrete phones of a single speaker. Other work has used a much larger vocabulary of continuous speech database of multiple speakers. Our system has been designed for specific applications such as control of machines using simple commands consisting of discrete utterances while other systems were developed for recognition of continuous speech. Nevertheless the

85% accuracies of our system is encouraging.

The authors suggest that one reason for the high accuracies of this system is that it is not only based on lip movement, but is based on the movement of the mouth, jaw and facial muscles. While lips are important articulators of speech, other parts of the mouth are also important, and this approach is closer to the accepted model of human visual speech perception.

The results demonstrate that a computationally inexpensive system which can easily be developed on a DSP chip can be used for such an application.

7 Conclusion

This paper reports on a voiceless speech recognition technique using video of the speaker's mouth that is computationally inexpensive and suitable for HCI applications. The proposed technique recognizes English consonants based on the dynamic speech information - facial movements of the speaker during phonation.

This paper adopts the MPEG-4 viseme model as the visual speech model to represent all the English consonants. An error rate of approximately 15% is obtained in classifying the consonants using this model. The misclassifications of the features can be attributed to the occlusion of speech articulators. Thus, non visible movements during the production of the consonants (such as movements of the tongue and vocal cords) are not represented in the visual speech features.

The results of our experiments suggest that the proposed technique is suitable in recognizing consonants using the information of the facial movements. The proposed system is easy to train for individual users and is designed for speaker-dependent speech-controlled applications. For future work, the authors intend to design a more suitable visual speech model for the consonants that accounts for the nonvisible articulators movements. Also, the authors intend to compare the performance of ANN with other classifiers such as Support Vector Machines (SVM) and Hidden Markov Models (HMM) to determine the optimum classifier for our application. Such a system could be used to drive computerized machinery in noisy environments. The system may also be used for helping disabled people to use a computer and for voice-less communication.

References

- Arjunan, S. P., Kumar, D. K., Yau, W. C. & Weghorn, H. (2006), Unspoken vowel recognition using facial electromyogram, *in* 'IEEE EMBC', New York.
- Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press.
- Bobick, A. F. & Davis, J. W. (2001), 'The recognition of human movement using temporal templates', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**, 257–267.
- Chen, T. (2001), 'Audiovisual speech processing', *IEEE Signal Processing Magazine* **18**, 9–21.
- Dikshit, P. & R.W.Schubert (1995), Electroglyph as an additional source of information in isolated word recognition, *in* 'Fourteenth Southern Biomedical Engineering Conference', LA, pp. 1–4.
- Freeman, A. & Skapura, M. (1991), *Neural Networks : Algorithms, Applications and Programming Techniques*, Addison-Wesley.
- Goecke, R. & Millar, J. B. (2003), Statistical analysis of the relationship between audio and video speech parameters for Australian English, *in* 'Proceedings of the ISCA Tutorial and Research Workshop on Auditory-Visual Speech Processing AVSP 2003', France, pp. 133–138.
- Gordan, M., Kotropoulos, C. & Pitas, I. (2002), Application of support vector machines classifiers to visual speech recognition, *in* 'International Conference on Image Processing', Vol. 3, Romania, pp. III–129 – III–132.
- Huang, K. Y. (2001), Neural network for robust recognition of seismic patterns, *in* 'IJCNN'01, Int Joint Conference on Neural Networks', Vol. 4, Washington, USA, pp. 2930–2935.
- Hazen, T. J. (2006), 'Visual model structures and synchrony constraints for audio-visual speech recognition', *IEEE Transactions on Audio, Speech and Language Processing* **14**(3), 1082–1089.
- Hazen, T. J., Saenko, K., La, C. H. & Glass, J. R. (2004), A segment-based audio visual speech recognizer : Data collection, development and initial experiments, *in* 'Int Conf on Multimodal Interfaces', State College, Pennsylvania, pp. 235–242.
- Kaplan, H., Bally, S. J. & Garretson, C. (1999), 'Speechreading: A way to improve understanding', pp. 14–16.
- Khontazad, A. & Hong, Y. H. (1990), 'Invariant image recognition by zernike moments', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**, 489–497.
- Kulkarni, A. D. (1994), *Artificial Neural Network for Image Understanding*, Van Nostrand Reinhold.
- Kumar, S. & Kumar, D. K. (2005), 'Visual hand gesture classification using wavelet transform and moment based features', *International Journal of Wavelets, Multiresolution and Information Processing (IJWMIP)* **3**(1), 79–102.
- Liang, L., Liu, X., Zhao, Y., Pi, X. & Nefian, A. V. (2002), Speaker independent audio-visual continuous speech recognition, *in* 'IEEE Int. Conf. on Multimedia and Expo', Vol. 2, Switzerland, pp. 25–28.
- Lippmann, R. P. (1997), 'Speech recognition by machines and humans', *J. Speech Communication* **22**, 1–15.
- Mallat, S. (1998), *A Wavelet Tour of Signal Processing*, Academic Press.
- Matthews, I., Cootes, T., Cox, S., Harvey, R. & Bangham, J. A. (1998), Lipreading using shape, shading and scale, *in* 'Proc. Auditory-Visual Speech Processing', Terrigal, Australia, pp. 73–78.
- Perez, J. F. G., Frangi, F. A., Solano, E. L. & Lukas, K. (2005), Lip reading for robust speech recognition on embedded devices, *in* 'ICASSP'05, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing', Vol. 1, Philadelphia, PA, USA, pp. 473–476.
- Petajan, E. D. (1984), Automatic lip-reading to enhance speech recognition, *in* 'GLOBECOM'84, IEEE Global Telecommunication Conference'.

- Potamianos, G., Neti, C., Gravier, G. & Senior, A. W. (2003), Recent advances in automatic recognition of audio-visual speech, *in* 'Proc. of IEEE', Vol. 91, pp. 1306–1326.
- Potamianos, G., Neti, C., Huang, J., Connell, J. H., Chu, S., Libal, V., Marcheret, E., Haas, N. & Jiang, J. (2004), Towards practical deployment of audio-visual speech recognition, *in* 'IEEE Int. Conf. on Acoustics, Speech, and Signal Processing', Vol. 3, Canada, pp. iii777–780.
- Saenko, K., Darrell, T. & Glass, J. (2004), Articulatory features for robust visual speech recognition, *in* 'ICMI'04', pp. 152–158.
- Simoncelli, E. P., Freeman, W. T., Adelson, E. H. & Heeger, D. J. (1992), 'Shiftable multiscale transform', *IEEE Transactions on Information Theory* **38**, 587–607.
- Soquet, A., Saerens, M. & Lecuit, V. (1999), Complementary cues for speech recognition, *in* '14th International Congress of Phonetic Sciences (ICPhs)', San Francisco, pp. 1645–1648.
- Stork, D. G. & Hennecke, M. E. (1996), Speechreading: An overview of image processing, feature extraction, sensory intergration and pattern recognition techniques, *in* '2nd International Conference on Automatic Face and Gesture Recognition (FG '96)', USA, pp. XVI–XXVI.
- Teague, M. R. (1980), 'Image analysis via the general theory of moments', *Journal of the Optical Society of America* **70**, 920–930.
- Teh, C. H. & Chin, R. T. (1988), 'On image analysis by the methods of moments', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **10**, 496–513.