Visual Speech Recognition and Utterance Segmentation based on Mouth Movement

Wai Chee Yau[1,2], Hans Weghorn [1]
[1]Information Technology
BA-University of Cooperative Education
Stuttgart, Germany
waichee@ieee.org, weghorn@ba-stuttgart.de

Dinesh Kant Kumar[2]
[2]School of Electrical and Computer Engineering
RMIT University
Melbourne, Australia
dinesh@rmit.edu.au

Abstract

This paper presents a vision-based approach to recognize speech without evaluating the acoustic signals. The proposed technique combines motion features and support vector machines (SVMs) to classify utterances. Segmentation of utterances is important in a visual speech recognition system. This research proposes a video segmentation method to detect the start and end frames of isolated utterances from an image sequence. Frames that correspond to 'speaking' and 'silence' phases are identified based on mouth movement information. The experimental results demonstrate that the proposed visual speech recognition technique yields high accuracy in a phoneme classification task. Potential applications of such a system are, e.g., human computer interface (HCI) for mobility-impaired users, lip-reading mobile phones, in-vehicle systems, and improvement of speech-based computer control in noisy environments.

1. Introduction

Speech technologies represent an important component in the development of next generation human computer interface (HCI). New HCI techniques emphasize on intelligent systems that can communicate with the users in a natural and flexible manner. The conventional human computer interfaces (HCI) such as mice and keyboards may not be suitable for people with limb disabilities. Users suffering from diseases or accidents such as strokes, amputations and amyotrophic lateral sclerosis may not be able to use their hands yet retaining the ability to speak. Speech-based systems are useful for such users to control the environment and to enhance their education and career opportunities. Nevertheless, speech recognition systems are not widely used as HCI due to the intrinsic sensitivity of such systems to variations in acoustic conditions. The performance of audio speech recognizers degrades when the sound signal strength is low, or in situations with high ambient noise levels.

Non audio sources can be used to identify utterances in an effort to overcome the limitations of voice-based speech systems. Options available are such as visual [19, 20], recording of vocal cord's movements [7] and recording of facial muscle activity [2]. This paper evaluates the use of images to identify speech. The visual signals are selected because the acquisitions of such data are non intrusive as opposed to methods that involves placement of sensors on users. The advantages of visual speech recognition are, e.g., not affected by audio noise and do not require users to make a sound. Such a system maybe useful for conveying confidential information and for military and defence applications.

Video recordings of a speaker contain information on the visible movement of the speech articulators such as lips, facial muscles, tongue and teeth. Research where audio and video inputs are combined to recognize large vocabulary, complex speech patterns are being reported in the literature [11, 20]. Without the voice signals, such systems have high error rate using only visual information [11, 20]. This suggest that the visual cues contain far less classification power for speech compared to audio data and hence it is to be expected to support only a small vocabulary of utterances.

Comprehensive reviews on speech recognition techniques can be found in [20, 22]. Visual features used in lip-reading systems can be divided into shape-based, intensity-based and motion-based. The shape-based features rely on the geometric shape of the mouth and lips. Such features usually can be represented by a small number of parameters. The first visual speech recognition system was proposed by Petajan [19] using shape-based features such as height, width and area of the mouth derived from the binary images. Researchers have reported on the use of artificial

0-7695-3067-2/07 $25.00 © 2007 IEEE
DOI 10.1109/DICTA.2007.53

markers on speaker's face to ease the extraction of the lip contours from the mouth images [1, 13]. The use of artificial markers is not suitable for practical speech-controlled applications. In [8], 3D coordinates of feature points such as lip corners are extracted from stereo images without using artificial markers. Lip contours can be extracted using active shape models (ASM) techniques that fit a statistical lip model into the video frames [17, 18]. Such top-down, model-based approaches are less sensitive to the view angle of the camera. An extension to the ASM technique is the active appearance model (AAM) approach that combines the shape model with a statistical model of the grey levels of the mouth region. The performance of AAM is demonstrated to outperform ASM in lip tracking [17]. Nevertheless, AAM and ASM techniques are sensitive to tracking error and modeling error.

Intensity-based features are obtained from the pixel intensity values of the image around the mouth area [20, 11]. The advantage of intensity-based systems is that accurate tracking and modeling of the lips are not required. 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, shaped-based features and lip contours [21]. Directly using all the pixels from the mouth images will result in very large size of feature vector. Feature extraction techniques such as Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Discrete Cosine Transforme (DCT) can be applied on the pixel values of 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 [17].

Features that represent the visual speech information through the different static poses of the mouth in individual frames can be viewed as static features. Features that directly utilize the dynamics of speech can be categorised as motion-based features. Few researchers have focused on motion-based features for visual speech recognition. Goldschen et. al. [9] demonstrates that dynamic visual features are most discriminative when comparing static and motion features. This paper proposes a visual speech recognition technique that utilizes a novel motion-based feature. These features are extracted by cascading multiple signal processing techniques including motion segmentation, Zernike moments and wavelet transform. This study examines a movement-based technique to detect the start and stop of utterances from video.

This research investigates the reliability of visual information in classifying a small set vocabulary of English phonemes. Earlier work by the authors has demonstrated a lip-reading technique that is insensitive to translation, rotation and scale changes of the mouth in the images using multilayer perceptron (MLP) neural network classifier [26]. Supervised neural network approach lends itself for identifying the separability of data even when the statistical properties and the type of separability (linear or nonlinear) are not known. While it may be an easy tool to implement as a first step, it may be suboptimal. To enhance the approach reported in [26], this paper proposes the use of support vector machines (SVMs) as speech classifier and evaluates the system on a larger vocabulary. One of the main advantages of SVMs is the ability such learning machines to achieve globally optimal solution. The use of SVMs in lip-reading system is reported in [10]. To model the temporal component of speech, the outputs of the SVM are integrated as nodes into a Viterbi lattice. This paper model the dynamic speech information in a different manner by using spatial-temporal templates named as motion history images (MHI). The proposed approach applies SVMs as discriminative classifier to classify the MHIs into phonemes.

This paper proposes a system where the camera is attached in place of the microphone to the commonly available head-sets to record mouth images. An advantage of this is that it is no longer required to identify the region of interest thereby reducing the computation required. This paper is organized as follows : Section 2 descibes the proposed feature extraction method that combined motion history images, wavelet transform and Zernike moments. Section 3 presents the design of the support vector machines (SVMs) speech classifier and Section 4 describes the methodology of the proposed visual speech recognition technique. Section 5 discusses the observations and findings from the experiments. Section 6 presents a new method for utterance segmentation based on movement information and Section 7 describes the conclusion of this paper and future work.

2. Motion feature

The first video processing step involved in the proposed approach is the segmentation of mouth motion. The facial movement of each utterance in the video file is represented using a 2D grayscale image - motion history image (MHI). MHI is a spatial-temporal template that shows where and when facial movements occur in the image sequence [3, 26].

MHI is generated using an accumulative image difference technique. The facial movement is segmented by detecting the changes between consecutive frames. Intensity values between successive frames of the video are subtracted to generate the difference of frames (DOFs). The DOFs are converted to binary images by thresholding the DOFs to obtain a change or no change classification. A fixed threshold value is determined heuristically through experimentation. The delimiters for the start and stop of the motion are manually inserted into the image sequence of every articulation. The binarised DOFs will have pixel value

1 at spatial coordinates where the intensity values between two consecutive frames are appreciably different. The intensity value of the MHI at pixel location (x, y) of tth frame is defined by

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

where N is the total number of frames of the video. $B_t(x,y)$ represents the binarised version of the DOF of frame t. In Eq. 1, $B_t(x,y)$ is multiplied with a linear ramp of time to implicitly encode the temporal information of the facial motions into the MHI. Each pixel value is a function of the temporal history of motion at that point from all the frames in the image sequence. By computing the MHI values for all the pixels coordinates (x, y) of the image sequence using Eq. 1 will produce a grayscale image (MHI) where the brightness of the pixels indicates the recency of motion in the image sequence [3].

MHI is used to segment the facial movement due to the ability of MHI to remove static elements and preserve the short duration facial movement in the video data. The MHI approach is computationally inexpensive and is insensitive to skin color due to the image subtraction process.

The speed of phonation of the speaker might vary for each repetition of the same phone. The variation in the speed of utterance results in the variation of the overall duration and there maybe variations in the micro phases of the utterances. The details of such variations are difficult to model due to the large 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 [0...1].

2.1. Wavelet transform

MHI is a view sensitive motion representation technique. MHI generated from the sequence of images is dependent on factors such as position, orientation and distance of the speaker's mouth from the camera. Also MHI is affected by small variations of the mouth movements while articulating the same phone. 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 [15]. 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. SWT decomposition of the MHI generates four images. The approximate image is the smoothed version of the MHI and carries the highest amount of information content among the four images. 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 reduce the size of the data. The pixel values of an LL image contain temporal information of the facial movement. Analyzing this information directly from the pixel values is difficult due to the large size of the data. For example, an LL image of size 240 x 240 has 57600 values. Further, the pixel values are sensitive to changes in scale, rotation and position of the mouth in the images.

2.2. Zernike moments

Image moments are feature descriptors that are concise, robust, and easy to compute and match. The proposed technique adopts Zernike moments as visual features to represent the SWT approximate image of the MHI. Zernike moments are selected by MPEG-7 as a robust region-based shape descriptor [12]. The main advantage of Zernike moments is the simple rotational property of the features[14].

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. V_{nl} is defined within a unit circle (i.e.: $x^2 + y^2 \leq 1$) given as follows:

$$V_{nl}(\rho,\theta) = R_{nl}(\rho)e^{-\hat{j}l\theta}; \hat{j} = \sqrt{-1} \quad (2)$$

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

Zernike moments are independent features due to the orthogonality of the Zernike polynomial V_{nl}[24]. $|l| \leq n$ and $(n - |l|)$ is even. Zernike moments Z_{nl} of order n and repetition l is given by

$$Z_{nl} = \left[\frac{n+1}{\pi}\right] \int_0^{2\pi} \int_0^{\infty} [V_{nl}(\rho,\theta)]\, f^*(\rho,\theta) d\rho d\theta \quad (3)$$

$f(\rho,\theta)$ 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 1 shows the square-to-circular transformation performed for the computation of the Zernike moments that transform the square image function, $f(x,y)$ to a circular image function $f(\rho,\theta)$. This transformation ensures the minimal lost of information as the entire square image is contained within the circular image.

To illustrate the rotational characteristics of Zernike moments, consider β as an angle that an image is rotated. The

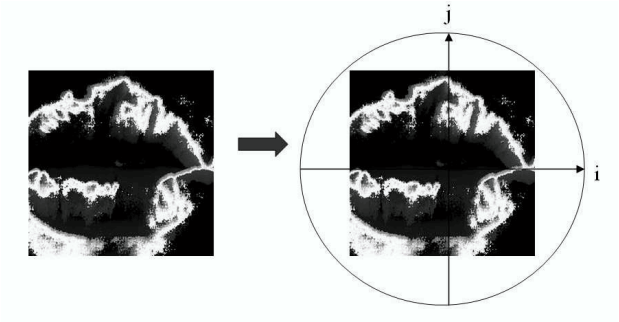

Figure 1. The square-to-circular transformation of the SWT approximation of MHI

resulting Zernike moments of the rotated image Z'_{nl} are given by

$$Z'_{nl} = Z_{nl}e^{-il\beta} \tag{4}$$

Z_{nl} is the Zernike moment of the original image. Eq. 4 demonstrates that rotation of an image results in a phase shift on the Zernike moments [23]. Hence, the absolute value of Zernike moments are rotation invariant [14] where

$$|Z'_{nl}| = |Z_{nl}| \tag{5}$$

This paper uses the absolute value of the Zernike moments, $|Z'_{nl}|$ as features of the SWT of MHI. The appropriate number of Zernike moment features to used is determined. 49 Zernike moments that comprise of 0^{th} order moments up to 12^{th} order moments have been found to yield reasonable performance and are used as features to represent the approximate image of the MHI.

3. Support vector machines classifier

This paper proposes the use of SVMs to classify the Zernike moments into visemes. SVMs are selected due to the ability of SVMs to determine the globally optimum hyperplanes to separate the different classes of the data. SVMs are learning machines that are trained based on the statistical learning theory [25]. The training of SVMs involved the minimizing the empirical error and complexity of the classifier simultaneously. Good generalization performance in SVMs is achieved by asserting bounds on the classification error and the capacity of the classifiers [4]. SVMs can be designed to classify linearly and non-linearly separable data.

In the case of non-linearly separable data, the data are not able to be separated using hyper plane in the original space. In SVM training, the data are projected to a higher-dimensional Hilbert space through nonlinear mapping. In the high-dimensional feature space, the data may be linearly separable using kernel functions [10]. Data that are not linearly separable in the original input space can be made to be linearly separable in the new feature space. SVM training can be carried out without knowing the nonlinear mapping explicitly. The commonly used non linear kernels are radial basis function kernels function (RBF) kernel and polynomial kernel. This paper implements nonlinear SVMs with polynomial order one kernel functions to classify the Zernike moment features.

4. Experiments

Experiments were conducted to test the proposed visual speech recognition technique. The experiments were approved by the RMIT University's Human Experiments Ethics Committee. A video speech database was recorded using a web camera in a typical office environment. This was done towards having a practical voiceless communication system using low resolution video data recorded in a realistic environment.

4.1. Vocabulary

Experiments were conducted to evaluate the performance of the system in classifying 14 English visemes. Visemes are the atomic units of visual movements associated with phonemes. This paper proposes the use of visemes to model visual speech because visemes can be concatenated to form words and sentences, thus providing the flexibility to increase the vocabulary of the system. The total number of visemes is much less than phonemes as speech is only partially visible [11]. The articulation of different speech sounds (such as /p/ and /b/) may be associated with identical facial movements. Each viseme may corresponds to more than one phoneme, resulting in a many-to-one mapping of phonemes-to-visemes. It is difficult to differentiate phonemes with identical facial motions based solely on the visual speech signals and hence other information from other sensory components is required to disambiguate these phonemes. Alternatively, language knowledge and context information may be used to differentiate such phonemes.

The number of visemes for English varies depending on factors such as the geographical location, culture, education background and age of the speakers. This paper adopts a viseme model established for facial animation applications by an international audiovisual object-based video representation standard known as MPEG-4. This model is selected to enable the proposed visual speech recognition system to be easily coupled with any MPEG-4 supported speech synthesis systems to form an interactive speech-based HCI. Based on the MPEG-4 viseme model shown in Table 1, the English phonemes can be grouped into 14

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

Phonemes	Vowel or Consonant
p, b, **m**	consonant
f,**v**	consonant
T , D	consonant
t , d	consonant
k, **g**	consonant
tS, dZ, S	consonant
s , z	consonant
n , l	consonant
r	consonant
A:	vowel
e	vowel
I	vowel
Q	vowel
U	vowel

visemes. The phonemes in bold fonts of each column are visemes tested in the experiments. Each visemes are repeated twenty times by a speaker. The camera focused on the mouth region of the speaker and 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. The video files were recorded and stored as true color (.AVI) files. The frame rate of the AVI files was thirty frames per second.

4.2. Feature extraction and classification

One MHI was generated from each utterance. SWT at level-1 using Haar wavelet was applied on the MHIs and the approximate image (LL) was used for analysis. Figure 2 shows an example of MHI of the fourteen visemes tested in the experiments. 49 Zernike moments have been used as features to represent the SWT approximate image of the MHI. 49 Zernike moments features were fed into the support vector machines (SVMs) classifier as input vectors. LIBSVM toolbox [6] was used in the experiment to design the c-SVMs. The one-vs-all multi-class SVM technique is adopted in the training of the SVMs. One SVM was created to learn each viseme. Three types of SVM kernel functions, (i) radial basis function, (ii) polynomial order one function and (iii) polynomial order three function were evaluated and compared in the experiments. The gamma parameter and the error term penalty parameter, c of the kernel function were optimized using ten-fold cross validation on the data. The performance of the SVM classifiers was evaluated using the leave-one-out method. 14 SVMs were trained with

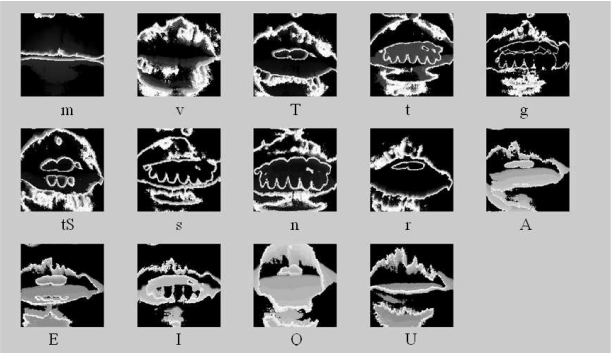

Figure 2. Motion history images (MHI) of fourteen vowels and consonants selected for experiments based on the MPEG-4 viseme model

266 training samples and were tested using the 14 remaining samples (1 sample from each viseme group). This process is repeated 20 times using different sets of train and test data. Three types of SVM kernel functions, (i) radial basis function, (ii) polynomial order one function and (iii) polynomial order three function were evaluated and compared in the experiments.

5. Results and Discussion

The classification accuracies of SVMs using different kernel functions are shown in Table 2. The performance of SVM classifier using polynomial order one kernel outperform the other SVM classifiers with radial basis function (RBF) kernel and polynomial order three kernel. The mean classification accuracy of the proposed lip-reading method is 91.4% using SVM classifier with first order polynomial kernel function. The promising results demonstrate that the proposed technique is suitable for identifying English visemes.

Comparing the results of the proposed SVM-based technique with the results of our earlier work using neural network [26] clearly indicates an improvement in classification accuracies. The neural network approach produces a lower recognition rate of 85% for a smaller vocabulary of nine consonants. The SVM classifier in our experiments yields a higher recognition rate for a larger set of vocabulary of 14 vowels and consonants. One of the possible reasons for the better performance of the SVM classifier is because SVM training seeks for an optimal solution as opposed to the training of neural networks that may be susceptible to local maxima. The higher accuracies of the SVMs classifier demonstrate the good generalization and capacity control of such learning machines. On the other hand, over training

Table 2. Recognition rates of the SVM Classifier using different kernel function

SVM Kernels	Recognition Rates (%)
1st order polynomial	91.4
3rd order polynomial	86.0
Radial basis function	86.8

or under training may occur in MLP neural networks and hence may results in misclassification of the visemes.

The result of SVM classifier is marginally higher than the classification accuracies obtained using continuous hidden Markov models (HMM) as reported in our previous work [27]. Approximately 3% improvement is achieved by using SVM classifier as compared to HMM which may be due to the smaller amount of training data required for training SVM.

6. Utterance segmentation

One of the challenges in recognizing speech based on video recordings is the segmentation of individual utterances, i.e., detecting the start frame and end frame of an utterance in a video clip containing multiple utterances. In audio-visual speech recognition techniques, speech segmentation is usually achieved through audio signals or by using transcribed video corpus (i.e. speech database that was manually annotated). In situations where audio signals are not available or highly contaminated by noise, video segmentation is required. This section describes a temporal segmentation framework to detect the start and end points of multiple isolated utterances in an image sequence.

6.1. Measure of mouth activity

The proposed method segment utterances from video clips based on mouth movement, without using the audio information. For isolated words (or phones) recognition task, a short pause is present in between two consecutive utterances. This pause periods generally consist of minimal mouth movement. When the speaker is pronouncing an utterance, relatively large mouth movements is produced. The level of mouth activity can be determined by computing MHIs for a number of consecutive frames. Figure 3 shows the average magnitude of the 49 Zernike moments used in the experiments for MHIs of utterances and MHIs of the 'pause' periods. Figure 3 indicates that the magnitude of Zernike moments corresponding to frames that contain utterances are much greater as compared to frames of the 'pause' or 'silence' period. To reduce the computation required, less number of Zernike moments can be used to

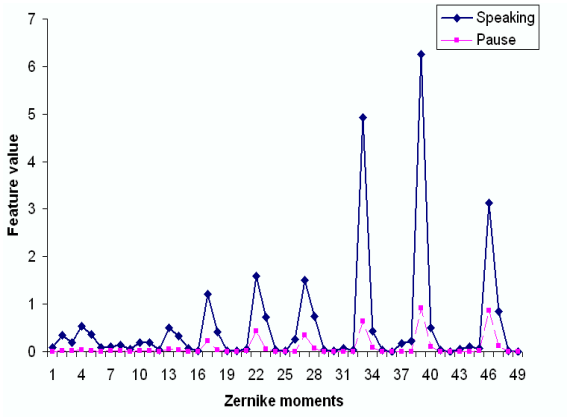

Figure 3. Fourty-nine Zernike moment of MHIs computed from 'speaking' frames versus 'pause' frames.

detect the mouth motion. The first six Zernike moments of the MHIs of utterances are at least an order of magnitude higher than the Zernike moments of MHIs computed from 'pause' period. This paper proposes a measure of the mouth movements (motion signals) based on a single parameter : the mean value of the first six Zernike moments ($Z_{00}, Z_{11}, Z_{20}, Z_{22}, Z_{31}, Z_{33}$).

6.2. Resolution of mouth motion signals

The 'resolution' (temporal details) of the motion can be adjusted by varying the number of consecutive frames for computation of MHIs. For example, MHIs calculated from a time window of 10 frames can represent movement information of approximately 333 ms of speed (for video files with a frame rate of 30 frames per second). This is achieved by computing one MHIs for every ten consecutive frames of the video data. If the time window of the MHIs are reduced to three frames, the resolution of the motion signals will be able to capture motion information of up to 100 ms. Figure 4 shows an example of motion signals of a video file computed from ten-frames MHIs and Figure 5 indicate the motion signals of the same video files computed from three-frames MHIs.

The smallest time-window possible to compute the MHI is two frames, where each MHI is equivalent to a difference of frame (DOF). Nevertheless, for 2-frame time window are found to be more susceptible to noise as compared to the 3-frame time window. 3-frame time window is selected to compute the MHI because it provides a good time resolution to capture the mouth movements for the utterance (for video data frame rate of 30 frames per second). The rate of speech differs based on factors related to individual, demographic, cultural, linguistic, psychological and physiologi-

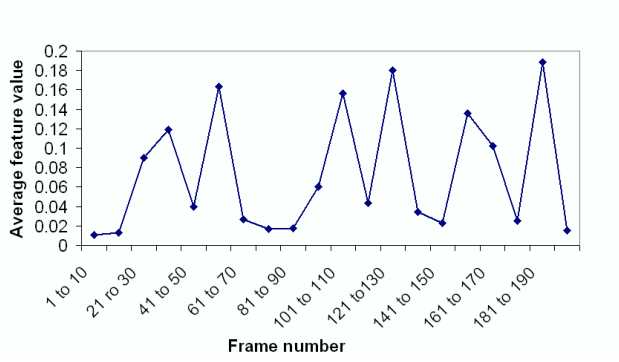

Figure 4. Averaged values of Zernike moments for MHIs computed from ten-frames time windows.

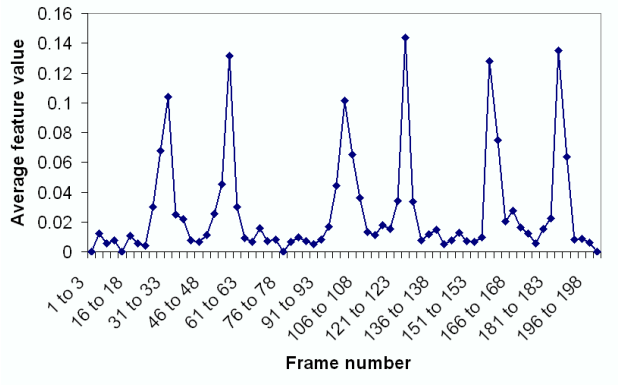

Figure 5. Averaged values of Zernike moments for MHIs computed from three-frames time windows.

cal [16]. An average rate of speech is 155 words per minute for native speakers of Australian English [5]. Based on this estimate, the mean period for a word is approximately 390 ms. Hence, the proposed segmentation technique using 3-frame time window with resolution of 100 ms is sufficient to capture the motion information of isolated words.

The mouth motion signals can be smoothed by using curve fitting techniques such as smoothing splines. Figure 6 shows the smoothed mouth motion signal (computed from MHIs of three-frames time window) for a video file that contains 3 repetitions of vowel /A/. Each pronunciation of the vowel is indicated by the shaded rectangular window. The first peak of the signal within each window represents the opening movement of the mouth and the second peak of each window indicates the closing movement of the mouth when pronouncing the vowel. Thus, based on the mouth activity, we can determine the start and stop of the individual utterances without using audio signals.

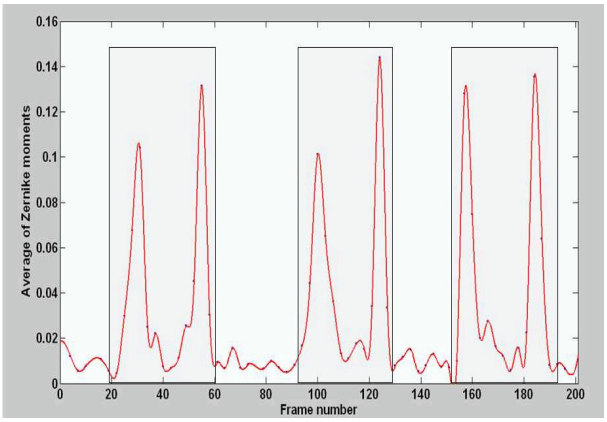

Figure 6. Mouth motion signals of 3 repetitions of vowel /A/. Each of the shaded windows represent one utterance.

The segmented utterance can be fed into the recognition sub-system described in previous sections to be identified as one of the command. The authors would like to point out that the proposed utterance segmentation approach are designed for isolated utterances and not continuous speech. Nevertheless, such a technique is useful for speech control for disabled users which may consist of isolated command words such as “on”, “off”, and digits. Such a technique can also be used for conveying confidential information such as pin codes and passwords to security systems and voice-dialing for mobile phones.

7. Conclusion

This paper describes a visual speech recognition method using video without evaluating audio signals. The proposed approach recognizes utterances from mouth images. The proposed technique identifies utterances based on mouth images using Zernike moments and support vector machines (SVMs). The promising results obtained demonstrate that the proposed technique can reliably identify English phonemes. The performance of SVM classifier is better as compared to neural network. A new framework for detecting the start and end of utterances from video data is proposed in this paper. Individual utterances are segmented based on the magnitude of mouth movement across consecutive frames.

For future work, the authors intend incorporate the segmentation framework into the proposed visual speech recognition system. Further, the investigation shall be extended from an English-spoken environment to other languages, e.g., German and Mandarin. Such a system may be

implemented for in-vehicle control and for helping disabled people to control computers. Future applications cover robotics and defense tasks involving voice-less communication.

References

[1] A. Adjoudani and C. Benoit. On the integration of auditory and visual parameters in an hmm-based asr. in speechreading by humans and machines: Models, systems, and application. In *Speechreading by Humans and Machines: Models, Systems, and Application*, pages 461–472, 1996.

[2] S. P. Arjunan, D. K. Kumar, W. C. Yau, and H. Weghorn. Unspoken vowel recognition using facial electromyogram. In *IEEE EMBC*. New York, 2006.

[3] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:257–267, 2001.

[4] C. Burges. *A tutorial on support vector machines for pattern recognition*, volume 2. Data Mining and Knowledge Discovery, 1998.

[5] C. S. C. Jones, L. Berry. Synthesized speech intelligibility and persuasion: Speech rate and non-native listeners. *Computer Speech and Language*, 21:641–651, 2007.

[6] C. C. Chang and C. J. Lin. Libsvm : a library for support vector machines. 2001,Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[7] P. Dikshit and R.W.Schubert. Electroglottograph as an additional source of information in isolated word recognition. In *Fourteenth Southern Biomedical Engineering Conference*, pages 1–4. LA, 1995.

[8] R. Goecke and J. B. Millar. 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*, pages 133–138. France, 2003.

[9] A. J. Goldschen, O. N. Garcia, and E. Petajan. Continuous optical automatic speech recognition by lipreading. In *28th Annual Asilomar Conf on Signal Systems and Computer*. Japan, 1994.

[10] M. Gordan, C. Kotropoulos, and I. Pitas. Application of support vector machines classifiers to visual speech recognition. In *International Conference on Image Processing*, volume 3, pages III–129 – III–132. Romania, 2002.

[11] T. J. Hazen. Visual model structures and synchrony constraints for audio-visual speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 14(3):1082–1089, 2006.

[12] S. Jeannin. Mpeg-7 visual part of experimentation model version 5.0. In *ISO/IEC JTC1/SC29/WG11/N3321*, 2000.

[13] M. N. Kaynak, Z. Qi, A. D. Cheok, K. Sengupta, and K. C. Chung. Audio-visual modeling for bimodal speech recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 34:564–570, 2001.

[14] A. Khontazad and Y. H. Hong. Invariant image recognition by zernike moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:489–497, 1990.

[15] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

[16] J. Y. Mark. Towards an integrated understanding of speaking rate in conversation.

[17] I. Matthews, T. Cootes, S. Cox, R. Harvey, and J. A. Bangham. Lipreading using shape, shading and scale. In *Proc. Auditory-Visual Speech Processing*, pages 73–78. Terrigal,Australia, 1998.

[18] J. F. G. Perez, F. A. Frangi, E. L. Solano, and K. Lukas. Lip reading for robust speech recognition on embedded devices. In *ICASSP'05, IEEE Int. Conf. on Acoustics , Speech, and Signal Processing*, volume 1, pages 473–476. Philadelphia, PA, USA, 2005.

[19] E. D. Petajan. Automatic lip-reading to enhance speech recognition. In *GLOBECOM'84, IEEE Global Telecommunication Conference*, 1984.

[20] G. Potamianos, C. Neti, G. Gravier, and A. W. Senior. Recent advances in automatic recognition of audio-visual speech. In *Proc. of IEEE*, volume 91, pages 1306–1326, 2003.

[21] G. Potamianos, C. Neti, J. Huang, J. H. Connell, S. Chu, V. Libal, E. Marcheret, N. Haas, and J. Jiang. Towards practical deployment of audio-visual speech recognition. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 3, pages iii777–780. Canada, 2004.

[22] D. G. Stork and M. E. Hennecke. Speechreading: An overview of image processing, feature extraction, sensory intergration and pattern recognition techiques. In *2nd International Conference on Automatic Face and Gesture Recognition (FG '96)*, pages XVI–XXVI. USA, 1996.

[23] M. R. Teague. Image analysis via the general theory of moments. *Journal of the Optical Society of America*, 70:920–930, 1980.

[24] C. H. Teh and R. T. Chin. On image analysis by the methods of moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:496–513, 1988.

[25] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[26] W. C. Yau, D. K. Kumar, and S. P. Arjunan. Visual recognition of speech consonants using facial movement features. In *Integrated Computer-Aided Engineering*, volume 14. IOS Press, 2007.

[27] W. C. Yau, D. K. Kumar, and H. Weghorn. Visual speech recognition using motion features and hidden markov models. In *12th Int. Conf. on Computer Analysis of Images and Patterns (CAIP 2007)*. Vienna, Austria, 27-29 August 2007.