Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Unspoken Vowel Recognition Using Facial Electromyogram

Sridhar P Arjunan, Dinesh K Kumar, Wai C Yau and Hans Weghorn

Abstract— **The paper aims to identify speech using the facial muscle activity without the audio signals. The paper presents an effective technique that measures the relative muscle activity of the articulatory muscles. Five English vowels were used as recognition variables. This paper reports using moving root mean square (RMS) of surface electromyogram (SEMG) of four facial muscles to segment the signal and identify the start and end of the utterance. The RMS of the signal between the start and end markers was integrated and normalised. This represented the relative muscle activity of the four muscles. These were classified using back propagation neural network to identify the speech. The technique was successfully used to classify 5 vowels into three classes and was not sensitive to the variation in speed and the style of speaking of the different subjects. The results also show that this technique was suitable for classifying the 5 vowels into 5 classes when trained for each of the subjects. It is suggested that such a technology may be used for the user to give simple unvoiced commands when trained for the specific user.**

I. INTRODUCTION

In our evolving technical world, it is important for human to have greater flexibility to interact and control our computers and thus our environment. Research on new methods of computer control has focused on three types of body functions: speech, bioelectrical activity and the use of mechanical sensors. Speech operated systems have the advantage that these provide the user with flexibility, and can be considered for any applications where natural language may be used. Such systems have the potential for making computer control effortless and natural. Further, due to the very dense information that can be coded in speech, speech based human computer interaction (HCI) can provide richness comparable to human to human interaction. In recent years, significant progress has been made in advancing speech recognition technology, making speech an effective modality in both telephony and multimodal human-machine interaction. However, speech recognition technology has three major shortcomings; (i) it is not suitable in noisy environments such as a vehicle or a factory (ii) it is not suitable for people with speech impairment disability, such as people after a stroke attack, and (iii) it is not suitable for giving discrete commands when there may be other people in the vicinity. This paper reports research to overcome these shortcomings, with the intent to develop a system that would identify the verbal command from the user without the need for the user to speak the command. The possible user of such systems would be people with disability, workers in noisy environments, and members of the defence forces. The identification of the speech with lip movement can be achieved using visual sensing, or sensing of the movement and shape using mechanical sensors[4] or by relating the movement and shape to the muscle activity[2], [3]. Each of these techniques has strengths and limitations. The video based technique is computationally expensive, requires a camera monitoring the lips and fixed to the user's head, and is sensitive to lighting conditions. The sensor based technique has the obvious disadvantage that it requires the user to have sensors fixed to the face, making the system not user friendly. The muscle monitoring systems have limitations of low reliability. This paper reports the use of recording muscle activity of the facial muscles to determine the unspoken command from the user.The paper has proposed techniques to overcome some of the limitations and it reports the development and testing of the use of using the relative contribution of the muscles when the spoken sounds are vowel-based phonemics. The paper also reports the analysis of the muscle activity with the corresponding sounds and has identified the possible limitations and applications of such a technique with respect to its reliability.

P A Sridhar, D K Kumar, W C Yau are with School of Electrical Engineering,RMIT University, GPO Box 2476V, Melbourne,Victoria 3001,Australia s3099587@student.rmit.edu.au

H Weghorn is with Information technology, BA-University of Cooperative Education, Rotebhlplatz 41, 70178 Stuttgart, Germany weghorn@ba-stuttgart.de

II. THEORY

The aim of this research is to classify the surface recordings of the facial muscle activity with speech. For the purpose of identifying the shape of the mouth and the muscle activity with speech, it is important to identify the anatomical details of speech production and it is convenient to divide the speech sounds into vowels and consonants. The consonants are relatively easy to define in terms of the shape and position of the vocal organs, but the vowels are less well defined and this may be explained because the tongue typically never touches another organ when making a vowel[8]. When considering the speech articulation, the shapes of the mouth during speaking vowels remain constant while during consonants the shapes of the mouth changes.

A. Facial movements and muscles related to Speech

The face can communicate a variety of information including subjective emotion, communitive intent, and cognitive appraisal. The facial musculature is a three dimensional assembly of small, pseudo- independently controlled muscular lips performing a variety of complex orfacial functions such as speech, mastication, swallowing and mediation of motion[7]. When using facial SEMG to determine the shape of the lips and the mouth, there is the issue of the choice of the muscles and the corresponding location of the electrodes.

1-4244-0033-3/06/$20.00 ©2006 IEEE.

There is also the difficulty of cross talk due to the overlap between the different muscles.Surface electromyogram (SEMG) is a gross indicator of the muscle activity and is used to identify force of muscle contraction, associated movement and posture[1]. Using an SEMG based system, Chan et al[2] demonstrated that the presence of speech information in facial myoelectric signals. Kumar et al[3] have demonstrated the use of SEMG to identify the unspoken sounds under controlled conditions.

The use of integral RMS of SEMG is useful in overcoming the issues of cross talk and the temporal difference between the activation of the different muscles that may be close to one set of electrodes. It is impractical to consider the entire facial muscles and record their electrical activity. In this study, only the following four facial muscles have been selected:*Zygomaticus Major*, *Depressor anguli oris*, *Masseter* and *Mentalis*[6]. With the intra and inter subject variation in the speed of speaking, and the length of each sound, it is difficult to determine a suitable window, and when the properties of the signal are time varying, this makes identifying suitable features for classification less robust. While each of these challenges are important, as a first step, this paper has considered the use of vowel based verbal commands only, where there is no change in the sound producing apparatus, the mouth cavity and the lips, and the nasal sounds can largely be ignored. In such a system, using moving RMS threshold, the temporal location of each activity can be identified.

B. Features of SEMG

SEMG is a complex and non-stationary signal. The strength of SEMG is a good measure of the strength of contraction of the muscle, and can be related to the movement and posture of the corresponding part of the body[1]. Root Mean Square(RMS) of SEMG is related to the number of active muscle fibers and the rate of activation, and is a good measure of the strength of the muscle activation.The issue regarding the use of SEMG to identify speech is the large variability of SEMG activity pattern associated with a phoneme of speech[1]. While it is relatively simple to identify the start and the end of the muscle activity related to the vowel, the muscle activity at the start and the end may often be much larger than the activity during the section when the mouth cavity shape is being kept constant, corresponding to the vowel. To overcome this issue, this research recommends the use of the integration of the RMS of SEMG from the start till the end of the utterance of the vowel. This paper reports the use of ratios of the area under the curve of RMS of SEMG from the different muscles to reduce the large inter-experimental variation.

III. METHODOLOGY

Experiments were conducted to identify and classify speech from facial EMG. The experiments were approved by the Human Experiments Ethics Committee of the University. Experiments were performed where electromyography (EMG) activity of suitable facial muscles was acquired from the subjects speaking 5 vowels.

A. EMG Recording and processing

Three male subjects participated in the experiment. The experiment used 4 channel EMG configurations as per the recommended EMG recording guidelines[6]. A four channel, portable, continuous recording MEGAWIN equipment (from MEGA Electronics, Finland) was used for this purpose. Raw signal sampled at 2000 samples/ second was recorded. Prior to the recording, the male participants were requested to shave their facial hair. The target sites were cleaned with alcohol wet swabs. Ag/AgCl electrodes (AMBU Blue sensors from MEDICOTEST, Denmark) were mounted on appropriate locations close to the selected facial muscles. The inter electrode distance was kept constant at 1cm for all the channels and the experiments. Controlled experiments were conducted where the subject was asked to speak the 5 English vowels (/a/,/e/,/i/,/o/,/u/).Each vowel was spoken separately such that there was a clear start and end of the utterance. During this utterance, facial SEMG from the muscles were recorded simultaneously. The recordings were visually observed, and the recordings with any artifacts typically due to loose electrodes or movement, were discarded. The experiment was repeated for ten times. A suitable resting time was given between each experiment. The participants were asked to vary their speaking speed and style to get a wide based training set.Fig.1 shows the raw EMG signal recorded from 4 channels (muscles). Example of the raw EMG signal recordings are plotted as a function of time (sample number) in Fig.1

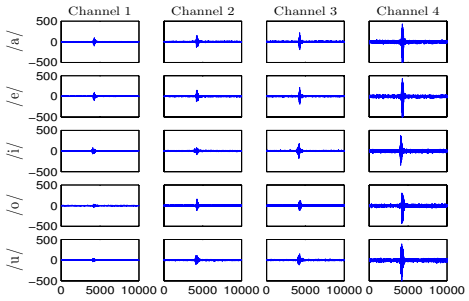

Fig. 1. Raw EMG signals recorded from Four muscles.

B. Data Analysis

The first step in the analysis of the data required identifying the temporal location of the muscle activity. Moving root mean square (MRMS) of the recorded signal was computed and thresholded against 1 sigma of the signal [10].The MRMS was computed using a moving window of 20 samples over the whole signal. Fig.2 is an example of the RMS plot of the recorded EMG signal. After identifying the start and the end of the muscle activity based on 1 sigma, these were confirmed visually. The RMS of the SEMG between the start and the end of the muscle activity was integrated for each of

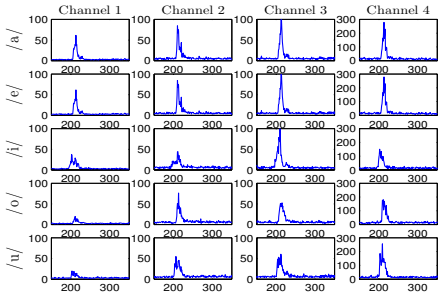

Fig. 2. RMS plot of the raw EMG signal.

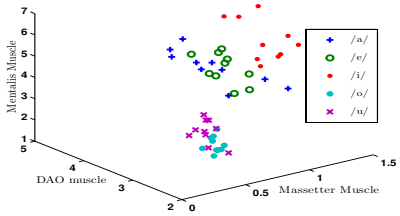

Fig. 4. Three dimensional plot of the normalised values of different muscles of different vowels

the channels. This resulted in one number representing the muscle activity for each channel for each vowel utterance. These were tabulated and all the channels were normalised with respect to channel 1 by taking a ratio of the respective integral with channel 1. This ratio is indicative of the relative strength of contraction of the different muscles and reduces the impact of inter-experiment variations. A demonstration of the computation of the integral of RMS of SEMG is shown in Fig.3. The paper reports the use of Durand's rule [9] for computing the integral of RMS of SEMG because it is applicable to problems in approximating areas and a straightforward family of numerical integration techniques.

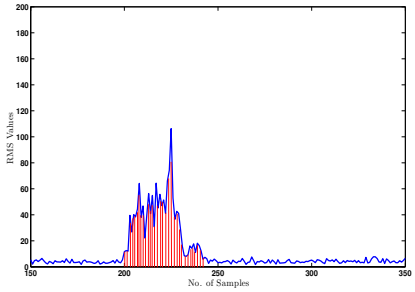

Fig. 3. A sample for finding the integral RMS of SEMG

C. Classification of Data

For each utterance of each vowel there were four numbers generated representing the total muscle activity by the four muscles. After normalisation with respect to the channel 1, this resulted in three set of numbers as the first channel was always one. As a first step, this data for each subject and for the five vowels and ten experiments were plotted on a three dimensional plot to visually identify any clusters. Data point from each of the vowels were given a distinct symbol and colour for ease of visual observation as shown in Fig.4. The data was then used to generate a dendrogram using Matlab based on average linkage method to identify the hierarchy of the clusters. The data from the ten experiments for each subject was divided into two separate sections; the training section and the test section. Each of these sections had data from five experiments. In the first part of the experiment, normalised integral RMS values of 5 recordings (for each vowel) for individual subject were used to train the artificial neural network (ANN) classifier with back propagation learning algorithm. In the second part of the experiment, the neural network was trained using the data clusters from Fig.4 by combining the data from vowel /a/, /e/ as one cluster, vowel /i/ as the second cluster and vowels /o/,/u/ as the third cluster for individual subject. In third part of the experiment, the neural network was trained using these data clusters as in the second experiment from one subject and tested with the other subjects. The architecture of the ANN consisted of two hidden layers and the 20 nodes for the two hidden layers were optimized iteratively during the training of the ANN. Sigmoid function was used as the threshold function and gradient descent and adaptive learning with momentum with a learning rate of 0.05 to reduce chances of local minima was used for training. In the testing section, the trained ANNs were used to classify the integral RMS values of 5 recordings of each vowel that were not used in the training of the ANN to test the performance of the proposed approach. This process was repeated for each of the subjects. The performance of these integral RMS values was evaluated in this experiment by comparing the accuracy in the classification during testing.

IV. RESULTS AND OBSERVATIONS

The results of the experiment report the performance of different subjects in classifying the integral RMS values of the 5 vowels. The three dimensional plot between the normalised area of the different muscle for different vowels is shown in Fig.4. The plot shows the different normalised values of SEMG integral for the different vowels forming clusters. It is evident from the plot that there are three different linearly separable clusters, or classes of vowels. Vowel /a/ and /e/ form one cluster, /i/ forms another cluster, while /o/ and /u/ form another cluster. This is further confirmed using hierarchical clustering shown in Fig.5. ANN are able to identify non-linear separations in clusters. The result of the use of these normalised values to train the ANN using data from individual subjects demonstrated easy convergence. Testing with the test data demonstrates an overall average accuracy of 80% of correct classification when testing and training data belong to the same subject. When testing the

TABLE I

CLASSIFICATION RESULTS FOR DIFFERENT SUBJECTS WHEN TRAINED AND TESTED INDIVIDUALLY

Vowel	Correctly Classified Vowels		
	Subject 1	Subject 2	Subject 3
/a/	3(60%)	2(40%)	2(40%)
/e/	2(40%)	3(60%)	3(60%)
/i/	4(80%)	4(80%)	4(80%)
/o/	2(40%)	1(20%)	2(40%)
/u/	3(60%)	4(80%)	4(80%)

TABLE II

CLASSIFICATION RESULTS FOR EACH SUBJECT WHEN TRAINED AND TESTED INDIVIDUALLY FOR THREE CLUSTERS OF VOWELS

Vowel	Correctly Classified Vowels		
	Subject 1	Subject 2	Subject 3
/a/,/e/	3(60%)	3(60%)	4(80%)
/i/	4(80%)	4(80%)	5(100%)
/o/,/u/	5(100%)	5(100%)	5(100%)

system with test data belonging to subjects different from the training data subject, the classification results showed lower accuracy. The classification results for each subject for five vowels when trained individually were analysed and tabulated in TableI I. It is observed that the classification accuracy for vowel /a/ and /e/ (60%) is marginally same for all subjects. The classification accuracy for vowel /i/ for all subjects (80%) is same and for vowel /o/ is low for the all subjects (40%) and the classification accuracy for vowel /u/ is 80%.

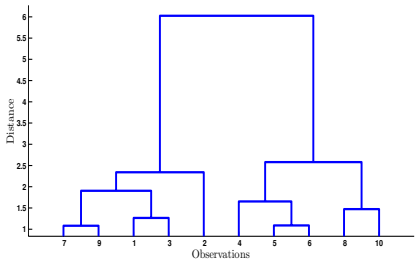

Fig. 5. Dendrogram of Observations using Hierarchial Clustering

The classification results for the subjects when trained data considering the three clusters for a subject and tested for a different subject is tabulated in Table III. The results indicate that the overall accuracy is poor. On closer observations, it is observed that while the system is unable to distinguish between /o/ and /u/, and also between /a/ and /e/. If the targets are reduced to three classes, with /o/ and /u/ being one class, /i/ being the second and /a/ and /e/ being the third class, the accuracy is very high, and the system identifies the utterance of the cluster /o/ & /u/ (accuracy 92%). This observation is similar from Fig.4. From the above, it is observed that the system is able to identify the differences between the five vowels for each individual, but when inter-subject variations are taken into account, the system is unable to distinguish between /a/ and /e/, and also between /o/ and /u/. The system

TABLE III

CLASSIFICATION RESULTS FOR EACH SUBJECT WHEN TRAINED INDIVIDUALLY FOR THREE CLUSTERS OF VOWELS AND TESTED WITH OTHER SUBJECTS

Vowel	Correctly Classified Vowels			
	Subject 1	Subject 2	Subject 3	Total
/a/,/e/	3	6	5	14(56%)
/i/	4	4	4	12(48%)
/o/,/u/	8	5	10	23(92%)

is sensitive to the styles of speaking of different people, and the inter-subject variation is high. This suggests that the system would be functional if trained for individual user.

V. DISCUSSION & CONCLUSION

This paper describes a voiceless vowel utterance recognition approach that is based on facial muscle contraction.The recognition accuracy is high when it is trained and tested for single userand it is poor when the system is used for testing the training network for all subjects. It should be pointed that this method at this stage is not being designed to provide the flexibility of regular conversation language, but for a limited dictionary only. The authors would also like to point out that this method is the system is easy to train for a new user. This method has to be enhanced for large set of data with many subjects in future.Speech generated facial electromyography signals could assist HCI by disambiguating the acoustic noise from multiple speakers and background noise.The results indicate that when used with multiple users, the system is able to classify only 3 groups of the 5 vowels. The system has been tested with a small set of phones, where the system has been successful, and appears to be robust despite variations in the speed of speaking. One possible application for such a system is for disabled user to give simple commands to a machine. Future possibilities include applications for telephony and defence.

REFERENCES

[1] J. V. Basmajian and C. J. Deluca, *Muscles Alive: Their Functions Revealed by Electromyography*,Fifth Edition,1985.

[2] D. C. Chan, K. Englehart, B. Hudgins, and D. F. Lovely,"A multi-expert speech recognition system using acoustic and myoelectric signals," *24th Annual Conference and the Annual Fall Meeting of the EMBS/BMES Conference*,2002.

[3] S. Kumar, D. K. Kumar, M. Alemu, and M. Burry, "EMG based voice recognition",*Intelligent Sensors, Sensor Networks and Information Processing Conference*, 2004.

[4] H.Manabe, A. Hiraiwa, and T. Sugimura,"Unvoiced speech recognition using SEMG - Mime Speech Recognition", *CHI*,2003.

[5] I.J.T. Veldhuizen, A.W.K. Gaillard and J. de Vries,"The influence of mental fatigue on facial EMG activity during a simulated workday",*Journal of Biological Psychology*,vol.63,2003.

[6] A.J.Fridlund and J.T.Cacioppo, "Guidelines for Human Electromyographic research",*Journal of Biological Psychology*,vol.23(5),1986.

[7] G. Lapatki, D. F. Stegeman, and I. E. Jonas,"A surface EMG electrode for the simultaneous observation of multiple facial muscles", *Journal of Neuroscience Methods*,vol. 123, pp. 117-128, 2003.

[8] Thomas .W. Parsons, *Voice and speech processing*,1986.

[9] Eric W. Weisstein,"Durand's Rule",From MathWorld-A Wolfram Web Resource. *http://mathworld.wolfram.com/DurandsRule.html*.

[10] David Freedman, Robert Pisani, and Roger Purves,"Statistics", *Third Edition*.