Facial SEMG for Speech Recognition Inter-Subject Variation

Sridhar P. Arjunan[1], Dinesh K. Kumar[1], Wai C. Yau[1] and Hans Weghorn[2]

[1] School of Electrical and Computer Engineering, RMIT Univesity,
GPO Box 2476V Melbourne, Victoria 3001, Australia

[2] Information technology, BA-University of Cooperative Education,
Rotebhlplatz 41,70178 Stuttgart, Germany

Abstract. The aim of this project is to identify speech using the facial muscle activity and without audio signals. The paper presents an effective technique that measures the relative muscle activity of the articulatory muscles. The paper has also tested the performance of this system for inter subject variation. Three 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, and the relative muscle activities of the four muscles were classified using back propagation neural network to identify the speech. The results show that this technique gives high recognition rate when used for each of the subjects. The results also indicate that the system accuracy drops when the network trained with one subject is tested with another subject. This suggests that there is a large inter-subject variation in the speaking style for similar sounds. The experiments also show that the system is easy to train for a new user. It is suggested that such a system is suitable for simple commands for human computer interface when it is trained for the user.

1 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 and development of new human computer interaction (HCI) techniques that enhance the flexibility and reliability for the user are important. 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 utilise a natural ability of the user. 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.

P. Arjunan S., K. Kumar D., C. Yau W. and Weghorn H. (2006).
Facial SEMG for Speech Recognition Inter-Subject Variation.
In *Proceedings of the 2nd International Workshop on Biosignal Processing and Classification*, pages 3-12
DOI: 10.5220/0001208600030012

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. Speech recognition systems have been built and deployed for numerous applications. The technology is not only improving at a steady pace, but is also becoming increasingly usable and useful. 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.

When we speak in noisy environments, or with people with hearing loss, the lip and facial movements often compensate the lack of quality audio. 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. The other difficulty of each of these systems is that these systems are user dependent and not suitable for different users. 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. The paper has evaluated such a system for its limitations, and reports the variation between different subjects. The paper 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 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 inter subject variations.

2 Theory

The aim of this research is to classify the surface recordings of the facial muscle activity with speech. For this purpose, the first step is to determine the role of the facial muscles in the production of speech. There are number of major speech production models that describe the mechanisms of speech productions.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.

2.1 Articulatory Phonetics

Articulatory phonetics considers the anatomical detail of the production speech sounds. This requires the description of speech sounds in terms of the position of the vocal organs. For this purpose, 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. The vowel is stationary, while the consonant is non-stationary.

2.2 Face Movement 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]. The parameterization used in speech is usually in terms of phonemes. A phoneme is a particular position of the mouth during a sound emission, and corresponds with specific sound properties. These phonemes in turn control the lower level parameters for the actual deformations. The required shape of the mouth and lips for the utterance of the phonemes is achieved by the controlled contraction of the facial muscles that is a result of the activity from the nervous system[8].

Surface electromyogram (SEMG) is the non-invasive recording of the muscle activity. It can be recorded from the surface using electrodes that are stuck to the skin and located close to the muscle to be studied. 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. There are number of challenges associated with the classification of muscle activity with respect to the associated movement and posture, such as the sensitivity of the location of electrodes, inter user variations, sensitivity of the system to variations in intrinsic factors such as skin conductance, and to external factors such as temperature, and electrode conditions. Veldhuizen et al[5] demonstrated the variation of facial EMG during a single day and has shown facial SEMG activity decreased during the workday and increased again in the evening.

One difficulty with speech identification using facial movement and shape is the temporal variation when the user is speaking complex time varying sounds. 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.The other difficulties also arise from the need for segmentation and the identification of the start and end of movement if the movement is complex. 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. Such a system would have limited vocabulary, and would not be very natural, but would be an important step in the evolution. In such a system, using moving RMS threshold, the temporal location of each activity can be identified. By having a stationary set of parameters defining the muscle activity for each spoken event, this also makes the system have very compact set of features, making it suitable for real time classification.

2.3 Facial Muscles for Speech

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. Face structure is more complex than the limbs, with large number of muscles with overlaps. It is thus difficult to identify the specific muscles that are responsible for specific facial actions and shapes. There is also the difficulty of cross talk due to the overlap between the different muscles. This is made more complex due to the temporal variation in the activation and deactivation of the different muscles. The use of integral of the 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. Due to the unknown aspect of the muscle groups that are activated to produce a sound, statistical distance based cluster analysis and back-propagation neural network has been used for classifying the integral of the RMS of the SEMG recordings. It is impractical to consider the entire facial muscles and record their electrical activity. In this study, only four facial muscles have been selected The *Zygomaticus Major* arises from the front surface of the zygomatic bone and merges with the muscles at the corner of the mouth. The *Depressor anguli oris* originates from the mandible and inserts skin at an angle of mouth and pulls corner of mouth downward.The *Masseter* originates from maxilla and zygomatic arch and inserts to ramus of mandible to elevate and protrude, assists in side-to-side movements mandible. The *Mentalis* originates from the mandible and inserts into the skin of the chin to elevate and protrude lower lip, pull skin into a pout[6].

2.4 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. The most commonly used feature to identify the strength of contraction of a muscle is the root mean square (RMS). RMS of SEMG is related to the number of active muscle fibres and the rate of activation, and is a good measure of the strength of the muscle activation, and thus the strength of the force of muscle contraction.

The preliminary study by Chan et al. has demonstrated the presence of speech information in facial EMG[2]. The timing of the activation of different groups of muscles is a central issue to identify the movement and shape of the mouth and lips. The issue regarding the use of SEMG to identify speech is the large variability of SEMG activity pattern associated with a phoneme of speech. A difference in the amount of motor unit

activity was observed in one and the same muscle when different words like p, b were spoken in the same context[1].

The vowels correspond to stationary muscle activity, the muscle activity pre and post the vowel is non-stationary. The other issue is the variation in the inter-subject because of variation in the speed and style of utterance of the vowel. 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. The temporal location of the start and the end of the activity is identifiable using moving window RMS.

Another shortcoming of the use of strength of SEMG is that it is dependent on the absolute of the magnitude of the recording, which can have large inter experimental variation. To overcome this shortcoming, this paper reports the use of ratios of the area under the curve of SEMG from the different muscles. By taking the ratio rather than the absolute value, the difficulty due the variation of the magnitude of SEMG between different experiments and between different individuals is overcome.

3 Methodology

Experiments were conducted to evaluate the performance of the proposed speech recognition from facial EMG by measuring the inter-subject variation on the system. The experiments were approved by the Human Experiments Ethics Committee of the University. Experiments were conducted where electromyography (EMG) activity of suitable facial muscles was acquired from the subjects speaking 3 vowels. As the muscle contraction is stationary during the utterance, root mean square values of each of the signals for the duration of the utterance was computed and used for further analysis.

3.1 EMG Recording and Processing

Three male subjects participated in the experiment. The experiment used 4 channel EMG configurations as per the recommended 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 muscles selected were the right side *Zygomaticus Major, Masseter & Mentalis* and left side *Depressor anguli oris*. 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. During this utterance, facial SEMG from the muscles was recorded. SEMG from Four channels were recorded simultaneously. The recordings were visually observed, and the recordings with any artefacts typically due to loose electrodes or movement, were discarded.

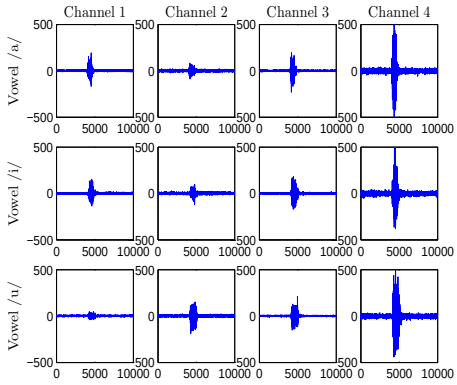


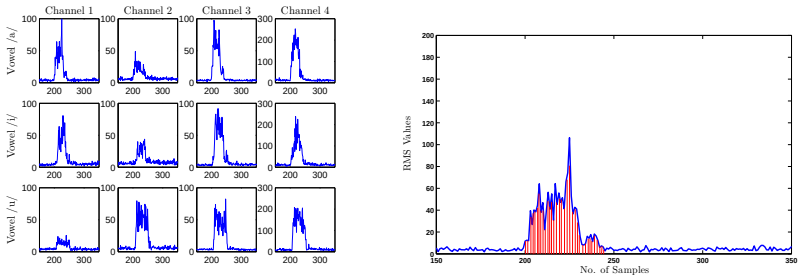
Fig. 1. Raw EMG signals recorded from Four muscles.

During the recordings, the subjects spoke the 3 English vowels (/a/,/i/,/u/). Each vowel was spoken separately such that there was a clear start and end of the utterance. 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

3.2 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(a) 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 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.2(b).

The paper reports the use of Durand's rule[9] for computing the integral of RMS of SEMG because it produces more accurate approximations and a straightforward family of numerical integration techniques. Durand rule states that 'Let the values of a function $f(x)$ be tabulated at points x_i equally spaced by $h = x_{i+1} - x_i$,so $f_1 = f(x_1), f_2 = f(x_2), \ldots f_n = f(x_n)$. Then Durand's rule approximating the integral of $f(x)$ is given by the Newton-Cotes-like formula'



(a) (b)

Fig. 2. (a)RMS plot of the EMG signal (b)A sample to find the integral RMS.

$$\int_{x_1}^{x_n} f(x)dx = h(2/5f_1 + 11/10f_2 + f_3 + \ldots + f_{n-2} + 11/10f_{n-1} + 2/5f_n) \quad (1)$$

The RMS values between the start and end of the muscle contraction was considered as x_1 to x_n, where h is the sample interval i.e., $x_{i+1} - x_i$.

3.3 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 three 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. This is shown in Fig.3.

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 ANN classifier with back propagation learning algorithm. In the second part of the experiment, the neural network was trained using the data from the three subjects simultaneously and tested similarly. 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 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 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 is repeated for different subjects. The performance of these integral RMS values was evaluated in this experiment by comparing the accuracy in the classification during testing.

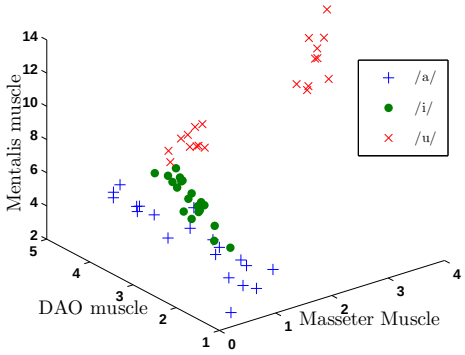


Fig. 3. Three dimensional plot of the normalised values of different muscles of different vowels.

4 Results and Observations

The results of the experiment report the performance of different subjects in classifying the integral RMS values of the 3 vowels. The three dimensional plot between the normalised area of the different muscle for different vowels is shown in Fig.3. The plot shows the different normalised values of vowels being separated from each other as clusters. It is evident from the plot that three different class of vowels form clusters that appear to be separate and distinct for each of the vowels. This is a clear indication that the three vowels are easy to separate using this technique. The result of the use of these normalised values to train the ANN using data from individual subjects demonstrated easy convergence. The results of testing the ANN to correctly classify the test data based on the weight matrix generated using the training data is tabulated in Table 1. The accuracy was computed based on the percentage of correct classified data points to the total number of data points. The results indicate an overall average accuracy of 80%, where it is noted that the overall classification of the integral RMS values of the EMG signal yields better recognition rate of vowels for 3 different subjects when it is trained individually.The classification results for each subject when trained individually were analysed to determine the inter-subject variation. It is observed that

- the classification accuracy for vowel /a/ for subject 1(80%) is marginally high when compared with the other subjects.
- the classification accuracy for vowel /i/ for subjects 2&3(80%) is equal and marginally high when compared with the subject 1.
- the classification accuracy for vowel /u/ is equal for the all subjects(100%).

The classification results for the subjects when trained for a subject and tested for a different is tabulated in Table 2. The results indicate that the overall accuracy is poor. On closer observations, it is observed that while the system is able to accurately identify the utterance of ‘/u/’ (accuracy 80%); the error for separating between ‘/a/’ and ‘/i/’ is poor.

Table 1. Classification results for different Subjects when trained and tested individually.

Vowel	Number of Utterances used for testing	Correctly Classified Vowels		
		Subject 1	Subject 2	Subject 3
/a/	5	4(80%)	3(60%)	3(60%)
/i/	5	3(60%)	4(80%)	4(80%)
/u/	5	5(100%)	5(100%)	5(100%)

This is also observable from the clustering in Fig.3. This suggests that while the system is able to identify the differences between the styles of speaking of different people at different times, the inter-subject variation is high. This suggests that the system would be functional if trained for individual subjects.

Table 2. Classification results for different subjects when trained with a subject and tested for other subjects.

Vowel	Number of Utterances used for testing	Correctly Classified Vowels			
		Subject 1	Subject 2	Subject 3	Total
/a/	25	4	1	1	6(24%)
/i/	25	3	2	1	6(24%)
/u/	25	5	9	6	20(80%)

5 Discussion

The results indicate that the proposed method provides better results for identifying the unvoiced vowels based on the movements of facial muscles. The recognition accuracy is high when it is trained and tested for single user. The accuracy of recognition is poor when the system is used for testing the training network for all subjects. This shows the large variations between subjects (inter-subject variation) because of different style and speed of speaking. This method has only been tested for limited vowels, because the muscle contraction during the utterance of vowels remain stationary The promising results obtained in the experiment indicate that this approach is suitable for classifying vowels based on the facial muscles movement of single user without regard to the speaking speed and style in different times. 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 in the system is easy to train for a new user. It is suggested that such a system is suitable for simple commands for human computer interface when it is trained for the user. This method has to be enhanced for large set of data with many subjects in

future. The authors have also used this method for checking the inter day variations of facial muscle activity for speech recognition.

6 Conclusion

This paper describes a voiceless speech recognition approach that is based on facial muscle contraction. The experiments indicate that the system is easy to train for a new user. Speech generated facial electromyography signals could assist HCI by disambiguating the acoustic noise from multiple speakers and background noise. This paper focused on classifying English vowels because pronunciation of vowels results in stationary muscle contraction as compared to consonants. The normalised integral RMS values of the facial EMG signals are used for analysis and classification of these values is performed by ANN. The results indicate that the system is reliable when trained for the individual user, while the inter-subject variation is large. The system has been tested with a very small set of phones, where the system has been successful, and appears to be robust despite variations in the speed of speaking. The inter-subject variation is high. The variation between different days and for a larger set of sounds is required to determine the possible applications. 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. Basmajian, J.V., Deluca, C.J.: Muscles Alive; Their Functions Revealed by Electromyography. Fifth Edition. (1985)
2. Chan, D.C., Englehart, K., Hudgins, B., Lovely, D. F.: A multi-expert speech recognition system using acoustic and myoelectric signals. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] EMBS/BMES Conference (2002).
3. Kumar, S., Kumar, D.K., Alemu, M., Burry, M.: EMG based voice recognition. Intelligent Sensors, Sensor Networks and Information Processing Conference (2004).
4. Manabe,H., Hiraiwa, A., Sugimura, T.: Unvoiced speech recognition using SEMG - Mime Speech Recognition. CHI (2003).
5. Veldhuizen, I.J.T., Gaillard, A.W.K., de Vries, J.: The influence of mental fatigue on facial EMG activity during a simulated workday. Vol. 63. Journal of Biological Psychology (2003).
6. Fridlund, A.J., Cacioppo, J.T.: Guidelines for Human Electromyographic research. Vol. 23(5). Journal of Psychophysiology (1986).
7. Lapatki, G., Stegeman,D. F., Jonas, I. E.: A surface EMG electrode for the simultaneous observation of multiple facial muscles. Vol 123. Journal of Neuroscience Methods (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, Roger Purves.: Statistics. Third Edition