Vowel Recognition using Facial Movement (SEMG) for Speech Control based HCI

Umesh Agnihotri
Research scholar, Electrical & Instrumentation Engineering Department, Sant Longowal Institute of Engineering & Technology, Longowal

Ajat Shatru Arora
Professor, Electrical & Instrumentation Engineering Department, Sant Longowal Institute of Engineering & Technology, Longowal

Atik Garg
B.E. Scholar,
Electrical & Instrumentation Engineering Department,
Sant Longowal Institute of Engineering & Technology,
Longowal

***Abstract*— This paper examines the use of facial muscle activity (Surface Electromyogram) to recognize speech based commands in English without any audio signals. The system is designed for applications based on speech control for Human Computer Interaction (HCI).The paper presents an effective technique that uses the facial muscle activity of the articulatory muscles and human factors for recognition. In these investigations, three English vowels were used as recognition variables. The moving root mean square (RMS) of surface electromyogram (SEMG) of four facial muscles is used to segment the signal and to identify the start and end of a silently spoken utterance. The relative muscle activity is computed by integrating and normalizing the RMS values of the signals between the detected start and end markers. The featured data of this is classified using a back propagation neural network to identify the voiceless speech. The experimental results show that this technique gives high recognition rate when used for each of the participants. The investigations also show that the system is easy to train for a new user.**

***Keywords*—**

I. INTRODUCTION

With the advancement of Technology, Automatic Speech Recognition (ASR) has used in various technologies and become popular in market in very short duration of time. Now markets are flooded with products which used speech Recognition as a standalone or as integrated technology like it is used in mobile phones to do various task from simple internet search to navigation purposes and even for speech to text dictation. And in now days these technologies are used for security of a device and in some place some robots are devolved to do the task on the basis of voice instruction. All these processes required careful analysis of voice sample and then execute that information properly. Despite of having so much advantage there are many major demerits of this technology. Firstly, Environmental Conditions affect a very great impact on this technology and degrade the performance in situation such as in Hotels, trains and other public places. Secondly you cannot keep your conversation private in many cases and thieves can also easily breach the voice recognition security New Voice analysis method and acoustic model adaption can compensate above three drawbacks at some extent, however the pervasive nature of mobiles become problem for this approach. Thirdly you have to speak loudly with a considerable manner for the faithful recognition of the voice but this might create annoyance to other people around you and moreover it also breaches your privacy and confidential information. Among many disadvantages, one of biggest drawback of this method is that people without vocal chords cannot use Automatic Speech Recognition and we need separate voice analysis procedure for different languages [1].

Due to these challenges and drawbacks traditional ASR researchers are witch to electromyography signals instead of acoustic signals [2]. This methods depends on the surface Electromyography (EMG) signal, where the potentials of the human articulatory muscles recorded by using several electrodes to follow the speech [3]. Fig.1 explains a basic setup for EMG-based silent speech interface [1]. After acquiring the EMG signals, they have to be converted to synthesized speech waveform or text information [4]-[5]. On the other hand, it can play an important role wherever humans interact with machines. Facial bioelectric signals have potential to mirror other bio signals such as EEG or EOG accompanying with SEMG [1]. In addition, Facial gestures could convey nonverbal expressions, which play an important role in interpersonal relations. There are many researches with different techniques which have been done in this area.

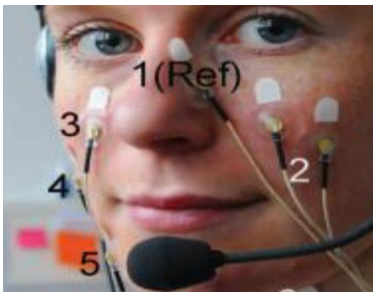

Fig.1: Electrodes positions, white number indicate bipolar derivation, blacks one indicate unipolar derivation [1]

Image processing is a popular and easy one with very low costs, feasible recognition (e.g. [2-4]). Surface electromyographic (EMG) signals-based facial gesture recognition has been considered lately (e.g. [5, 6], [24]). Using bio-sensors mounted on facial muscles. This method has some privileges over other gesture recognition methods. It is strong against many environmental circumstances which are difficult to overcome by other methods of gesture recognition [8]. One of the most important issues is the number of bioelectric data channels (the number of electrodes). This paper reports research to over-come these shortcomings, with the intent to develop system that would identify the verbal command from the user without the need for the user to speak the command. This paper reports the use of recording muscle activity of the facial muscles to determine the unspoken command from the user. Earlier work reported by the authors has demonstrated the use of multichannel surface Electromyogram (SEMG) to identify the unspoken vowel based on the normalized integral values of SEMG during the utterance.

II. RELATED WORK

EMG signal has been used in many researches for speech recognition since 1980s. The first two studies in this direction were done by Sugie and Tsunda in 1985 and Morse eta al in 1986[7]-[8]. In the first one, Sugie and Tsunda used three electrodes to acquire the EMG signal. The acquired signal has to be passed to three channels EMG signal synthesizer system. This system has been developed to recognize Japanese vowel in real time. The channel output is "1" if it is in active state and "0" if it is in inactive state (no movement). The output of these channels was fed to a finite automaton for vowel discrimination. The output of this automaton was used to derive speech synthesizer for vowel production [7]. In the second paper, Morse et al provided a speech recognition system to discriminate among ten spoken words in English [8]. This paper was the extension of the work that was done by Morse and O'Brien when they examined speech information from neck and head muscles activity to discriminate between two words [9]. These earlier studies gave very encouraging results but after increasing the number of vocabulary the recognition accuracy decreased dramatically. Many researchers have been tried to improve the recognition accuracy for large numbers of vocabulary, Chan et al have achieved first an accuracy rate of 93% on a ten English words [10]. Before this study, Morse et al had achieved 60% of accuracy by using back propagation neural network instead of time domain analysis [11]. In 1991, Bahl et al provided a model to simulate the variation of the pronunciations of the words in acoustic speech recognition which is one of the biggest challenges in EMG-based speech recognition nowadays [12]. This process has been done manually in the past, so Bahl etal. used their model to do it automatically. In 1998, Kain and Macon developed a new spectral conversion algorithm [13] that uses local linear transformation based on Gaussian Mixture Model (GMM). This study has explained the importance of the pitch in the speaker identification EMG signal has been used for silent speech recognition. Jorgensen et al recorded the EMG signal using two electrodes at the larynx and sublingual areas below the jaw [14]. Their signal has to be filtered to remove the noise and artifacts from it before getting its features using complex dual quad tree wavelet transform. This process has been tested on six basic words; stop, go, right, left, alpha, and omega. The results showed that the EMG signal is richen of speech information. Hence this information can be used as a way to recognize the speech. In 2005, Maier-Hain et al studied about repositioning of the electrodes and its effect on recognition accuracy [15]. The accuracy they got was 97.3% for the within session case. This percent reduced significantly if the electrode were repositioned to 76.2%. By using normalization and adaptation methods, the recognition accuracy can be increased back to 87.1%. The results also showed that using two electrodes is crucial while using more than five electrodes will not affect much. They suggested using data from different sessions to improve the adaptation which leads to enhance the recognition accuracy. On the other hand, some researchers have tried to use other media from speech [16]. The most interesting thing is that the speech recognition based on EMG signal is not limited to certain language. Besides the English which is the dominant one, there are other studies have been done in Japanese, Chinese, and other languages. The first paper studied the recognition of Chinese digits from zero to ten was presented by Jia et al [17]. In this study, wavelet transform coefficients, auto regressive model were calculated as the feature of the myoelectric signal. Besides this paper, there is another study of recognition six Chinese vowels that was done by Jia et al [18]. In 2006, several researchers suggested that the phonemes could be used as a modeling unit for speech recognition on EMG signal instead of the whole word [4], [19]. In spite of the great progress that has been made over the last few years, EMG-based speech recognition still faced several challenges in the large numbers of vocabulary recognition. The main challenges are effect of speaking modes and the repositioning of electrodes. These factors have pushed many researchers around the world to alternate from session-dependent to session-independent system. The system has been tested for both cases, session-dependent and session independent. The results he got showed that the dependent model gives better words recognition rate which is 85.7% approximately [20].

III. EMG SIGNAL ACQUISITION

The EMG signals vary among speakers, and they also change with the same speaker across different sessions. Based on this fact, the performance across speakers and sessions might be unstable. To avoid the problem of instability, some researchers have used one speaker and one session data during their study [32]. To get the best EMG signal to noise ratio, recording has been done in a very quiet room.

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.

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 (Fridlund 1986). The location of these muscles is shown in Figure below.

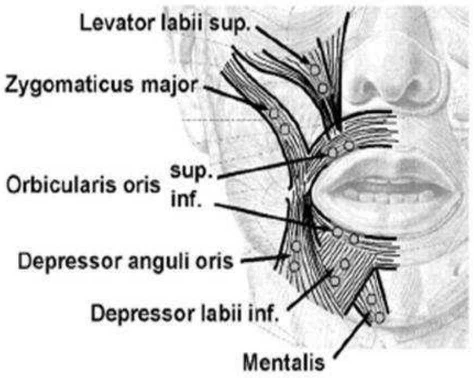

Fig.2: Facial muscles

The drawbacks are such a system would have limited vocabulary, and would not be very natural, but would be an important step in the evolution. The electrodes were arranged on the distance of 1cm with the help of flexible ruler. In [14], the signals were collected using two pairs of electrodes located at the right and left area of throat near the chin cleft and 1-0.5 centimeter from the left and right side of the larynx. A notch filter of 60 Hz was used to remove the power line interference. Jia et al [18] used electrodes to get the signal after deep study of electrodes positions from Chan et al and Hiroyuki Manabe [10], [34]. It has to be mentioned that in each recording process a reference electrode has to be available but placed at except interactive active muscle. After acquiring the EMG signal, it has to be passed through preprocessing and processing steps before getting its features.

IV. PREPROCESSING

One of the biggest challenges in EMG-based speech recognition is the noise and artifacts in the recorded signal [35]. To overcome this problem a filtering process should be done on the received signal before getting features and information contained in it. The main goal of the preprocessing step is increasing the signal to noise ratio by reducing the noise. Some researchers use filters such as low pass filter, high pass filter, and low pass filter. Beside these filters, notch filter has been used to remove the power line interference [21]. Jia et al [18] used wavelet transform to remove this effect. According to them, wavelet transform is not only a very promising technique for time –frequency analysis but also a noise reduction method. Proper skin preparation is important to get a good signal and avoid artifacts. Before electrode placement, the selected areas which are proper for signal recording must be cleaned from any dust, sweat or fat layers to reduce the effect of motion artifacts [6], [24]. Conductive electrode paste or cream is used on the center of electrodes (grey area only) before applying them to the skin. Two pairs of rounded pre-gelled Ag/AgCl electrodes are placed on the volunteer’s facial muscles in a different configuration to provide the highest amplitude signals [5].

V. SIGNAL PROCESSING

A. *Filtering*

It is well-known that the electrode’s configuration has a band pass filtering effect in the EMG signal spectral range. So to overcome this problem, a filtering process has to be done. Instead of using band pass filter to remove these effects, a combination of low pass and high pass filters are usually used to avoid the aliasing problem. Besides these filter, a 50/60 Hz notch filter is used often to remove power line interferences as we explained before. The filtering process is usually considered as preprocessing steps.

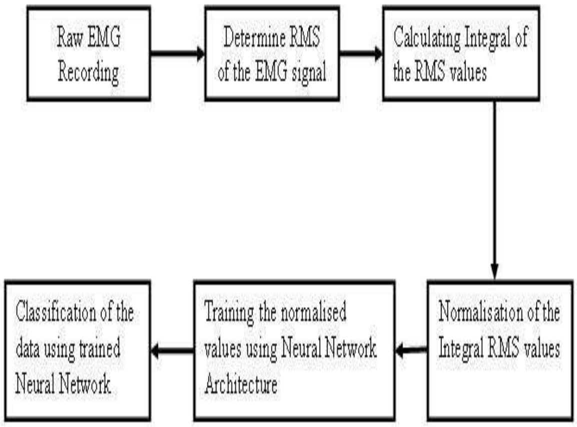
Raw EMG Recording
Determine RMS of the EMG signal
Calculating Integral of the RMS values
Normalisation of the Integral RMS values
Training the normalised values using Neural Network Architecture
Classification of the data using trained Neural Network

Fig.3: A simplified block diagram of methodology

B. *Normalization*

The EMG signal is very sensitive to the changes in the electrodes positions, and temperature or tissue properties in our bodies. Hence to make a comparison of possible amplitudes, it is very important to apply a normalization process at each recording in order to compensate these changes. Isometric Maximal Voluntary Contraction (MVC) is a widely used method for normalization purposes [38].

VI. FEATURE EXTRACTION

SEMG is a complex and non-stationary signals whose power is an excellent measure for the strength of the muscles contraction. In addition, it can be related to the movement and position of the corresponding part of the body. RMS of SEMG speech signals is associated with the number of active muscle fibers and the rate of activation; it’s a good measure for the strength of the muscle activation, and consequently for the strength of the muscle contraction. 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.

VII. CLASSIFICATION OF SEMG DATA

The first step in classification of data was to determine if this data was separable. After confirming this, the next step undertaken was to determine whether the data is linearly separable. To determine whether the data is separable, supervised neural network approach was used. The advantage of using such a neural network is that neural networks can be applied without the assumption for linear separation of the data. For this purpose, the data from the experiments for each participant was divided into two equal groups training and test data. The ANN consisted of two hidden layers with 20 nodes in both layers. Sigmoid function was used as the threshold decision. This entire process was repeated for each of the participants. The performance of these integral RMS values was evaluated in this experiment by comparing the accuracy in the classification during testing. The accuracy was computed based on the percentage of correctly classified data points to the total number of data points in the class. The advantage of ANN approach is that ANN is easy to be trained by a user to configure the system for the individual.

VIII. RESULTS AND OBSERVATIONS

Results indicate an overall average accuracy of 84.9%, 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 participants, when it is trained individually.

Table-1:

vowel	Subject					avg %
	1	2	3	4	5	
A	100	78.9	94	94.4	75	88.46
E	65	77.1	84	77.8	93.3	73.44
O	91.32	76.9	95.4	93	100	93.06

The results indicate that this technique can be used for the classification of vowels and suggests that the system is able to identify the differences between the styles of speaking of different people at different times for different languages. The recognition accuracy is high, when it is trained and tested for a dedicate user. Hence, such a system could be used by any individual user as a reliable human computer interface (HCI). This method has only been tested for limited vowels. Vowels were the first to be considered because the muscle contraction during the utterance of vowels remains stationary. The promising results obtained in the experiment indicate that this approach based on the facial muscles movement represents a suitable, reliable method for classifying vowels of single user without regard to the speaking speed and style in different times for different languages. The results furthermore suggest that such a system is suitable and reliable 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.

IX. CONCLUSION

This paper describes a silent vowel based speech identification approach that is based on measuring the facial muscle contraction using non-invasive SEMG. The experiments indicate that the system is easy to train for a new user. The presented investigation focused on classifying English vowels, because pronunciation of vowels results in stationary muscle contraction as compared to consonants. The results indicate that the system is reliable when trained for the individual user. One possible application for such a system is for disabled user to give simple commands to a machine which is a good and typical application of HCI. Future possibilities include applications for telephony, defense problems and improvement of speech-based computer control in noisy environments.

REFERENCES

[1] M. Janke, M. Wand, K. Nakamura, and T. Shultz, “Further investigation on EMG-to-speech conversion,” in Proceedings of IEEE (ICASSP), Kyoto, Japan, pp: 365–368, March 25-30 ,2012.

[2] B. Denby, T. Schultz, K. Honda, T. Hueber, and J. Gilbert, “Silent speech interfaces”, Speech Communication, vol. 52, no. 4, pp: 270–287, 2010.

[3] S. C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, “Towards continuous speech recognition using surface electromyography”, in Proceedings of Inter speech 2006, Pittsburgh, Pennsylvania, pp: 573–576, September 1, 2006.

[4] A. R. Tot h, M. Wand, and T. Schultz, “Synthesizing speech from electromyography using voice transformation techniques”, in Proceedings of Inter speech 2009, Brighton, United Kingdom, pp: 652–655, 2009.

[5] M. Wand, A. Toth, S.C. Jou, and T. Schultz, “Impact of different speaking modes on EMG-based speech recognition”, in Proceedings of Inter speech, Brighton, United Kingdom, pp: 648–651, 2009.

[6] M. S. Morse, S. H. Day, B. Trull, and H. Morse, “Use of myoelectric signal to recognize speech,” in Proceedings of 11th Annual Conference of the IEEE Engineering in Medicine and Biology Society, vol. 6 pp: 1793–1794, 1989.

[7] M. S. Morse, N. Y. Gopalan, and M. Wright ,“Speech recognition using myoelectric signals with neural networks”, in Proceedings of 13th Annual Conference of the IEEE Engineering in Medicine and Biology Society, vol. 13, pp: 1877–1878, 1991.

[8] M. S. Morse, and E. M. O'Brien, “Research summary of a scheme to ascertain the availability of speech information in the myoelectric signals of neck and head muscle using surface electrodes”, Computers in Biology and Medicine, vol. 16, no. 6, pp: 399-410, 1986.

[9] A. Chan, K. Englehart, B. Hudgins, and D. Lovely, “Myoelectric signal to augment speech recognition,” Medical and Biological Engineering and Computing, vol. 39, no. 4, pp: 500-506, 2001.

[10] A. Kain and M. W. Macon, “Spectral voice conversion for text - to-speech synthesis”, in Proceedings of IEEE (ICASSP), Beaverton, Oregon, May 12-15, 1998.

[11] C. Jorgensen, D. D. Lee, and S. Agabon, “Sub auditory speech recognition based on EMG signals”, in Proceedings of International Joint Conference on Neural Network, Moffett Field, California, July 20-24, 2003.

[12] L. Maier-Hein, F. Metze, T. Shultz and A. Waibel, “Session independent non-audible speech recognition using surface electromyography”, IEEE Workshop on Automatic Speech Recognition and Understanding, pp: 331-336, 2005.

[13] X. Jia, X. Wang, J. Li, D. Yang, and Y. Song, “Unvoiced Chinese digital recognition based on facial myoelectric signal”, in Proceedings of IEEE International Conference on Communication, Circuits and Systems, Shenyang, China, June 25-28, 2006.

[14] M. Walliczek, F. Kraft, S. C. Jou, T. Shultz, and A. Waible, “Sub-word unit based non-audible speech recognition using surface electromyography”, in Proceedings of Inter speech 2006, Pittsburgh, Pennsylvania, September 1, 2006.

[15] Q. Zhou, N. Jiang, K. Englehart, and B. Hudgins, “Improved phoneme-based myoelectric speech recognition”, IEEE Transaction on Biomedical Engineering, vol. 56, no. 8, August 2009.

[16] M. Wand and T. Shultz, “Session-independent EMG-based speech recognition”, in Proceedings of International Conference on Bio-Inspired Systems and Signal Processing, pp: 295-300, 2011.

[17] M. Wand and T. Shultz, “Analysis of phone confusion in EMG based speech recognition”, in Proceedings of IEEE (ICASSP), Prague, Cheek Republic, pp: 757-760, May 22-27, 2011.

[18] M. Janke, M. Wand, and T. Schultz, “Spectral energy mapping for EMG-based recognition of silent speech”, in Proceedings of First International Workshop on Bio-inspired Human-Machine Interfaces and Healthcare Applications, 2010.

[19] M. Wand and T. Shultz, “Adaptive speech recognition based on surface electromyography," Proceedings of Bio signals, Porto, Portugal, pp: 155-162, 2009.

[20] S. C. Jou, T. Shultz, and A. Waible, “Continuous electromyographic speech recognition with a multi-stream decoding architecture”, in Proceedings of IEEE (ICASSP), Honolulu, Hawaii, pp: 401-404, 2007.

[21] E. J. Scheme, “Myoelectric signal classification for phoneme based speech recognition”, M.Sc thesis, University of New Brunswick, Canada, 2005.

[22] H. Manabe, A. Hiraiwa, and T. Sugimura, "Unvoiced speech recognition using EMG-mime speech recognition”, in Proceedings of the Conference on Human Factors in Computing Systems, Ft. Lauderdale, Florida, 2003.

[23] K. Das, S. Osechinskiy, and Z. Nenadic, “A class wise PCA-based recognition of neural data for brain-computer interfaces”, in Proceedings of IEEE 29th Annual International Conference in EMBS, pp: 6519-6522, 2007.

[24] K. Englehart, B. Hudgins, P. Parker, and M. Stevenson, “Classification of the myoelectric signal using time-frequency based representations”, Medical Engineering and Physics, no. 21, pp: 431-438, 1999.