SILENT BILINGUAL VOWEL RECOGNITION
Using fSEMG for HCI based Speech Commands

Sridhar Poosapadi Arjunan
School of Electrical and Computer Engineering, RMIT University,GPO Box 2476V, Melbourne, VIC 3001, Australia

Hans Weghorn
Information technology, BA-University of Cooperative Education, Rotebuehlplatz 41, 70178 Stuttgart, Germany

Dinesh Kant Kumar, Wai Chee Yau
School of Electrical and Computer Engineering, RMIT University,GPO Box 2476V, Melbourne, VIC 3001, Australia

Keywords: HCI, Speech Command, Facial Surface Electromyogram, Artificial Neural Network, Bilingual variation.

Abstract: This research examines the use of fSEMG (facial Surface Electromyogram) to recognise speech commands in English and German language without evaluating any voice signals. The system is designed for applications based on speech commands for Human Computer Interaction (HCI). An effective technique is presented, which uses the facial muscle activity of the articulatory muscles and human factors for silent vowel recognition. The difference in the speed and style of speaking varies between experiments, and this variation appears to be more pronounced when people are speaking a different language other than their native language. This investigation reports measuring the relative activity of the articulatory muscles for recognition of silent vowels of German (native) and English (foreign) languages. In this analysis, three English vowels and three German 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 normalising the RMS values of the signals between the detected start and end markers. The output vector of this is classified using a back propagation neural network to identify the voiceless speech. The cross-validation was performed to test the reliability of the classification. The data is also tested using K-means clustering technique to determine the linearity of separation of the data. The experimental results show that this technique yields high recognition rate when used for all participants in both languages. The results also show that the system is easy to train for a new user and suggest that such a system works reliably for simple vowel based commands for human computer interface when it is trained for a user, who can speak one or more languages and for people who have speech disability.

1 INTRODUCTION

In this advancing world of technology, there are many developments being made in the field of computing. Research and development of new human computer interaction (HCI) techniques that enhance the flexibility and reliability for the user are important. The most fundamental applications of affective computing would be human-computer interaction, in which the computer is able to detect and track commands coming from human users, and to handle communication based on this knowledge. Research on new methods of computer control has focused on various types of body functions like speech, emotions, bioelectrical activity, facial expressions, etc. The expression of emotions plays an important part in human interaction. Most of the facial movements result from either speech or the display of emotions; each of these has its own complexity (Ursula and Pierre, 1998).

Speech operated systems have the advantage that these provide flexibility, and can be considered

Poosapadi Arjunan S., Weghorn H., Kant Kumar D. and Chee Yau W. (2007).
SILENT BILINGUAL VOWEL RECOGNITION - Using fSEMG for HCI based Speech Commands.
In *Proceedings of the Ninth International Conference on Enterprise Information Systems - HCI*, pages 68-75
DOI: 10.5220/0002365400680075

for any applications where natural language may be used. Such systems utilise a natural ability of the human user, and therefore have the potential for making computer control effortless and natural. Furthermore, 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 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 using voice signals has three major shortcomings - it is not suitable in noisy environments such as vehicles or factories, not applicable for people with speech impairment disability such as people after a stroke attack, and it is not applicable for giving discrete commands when there may be other people talking in vicinity.

This research reports how to overcome these shortcomings with a voice recognition approach, which identifies silent vowel-based verbal commands without the need to sense the voice sound output of the speaker. Possible users of such a system 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 deficiencies, the lip and facial movements often compensate the lack of audio quality.

The identification of speech by evaluating lip movements can be achieved using visual sensing, or tracking the movement and shape using mechanical sensors (Manabe et al., 2003), or by relating the movement and shape to facial muscle activity (Chan et al., 2002; Kumar et al., 2004). Each of these techniques has strengths and limitations. The video based technique is computationally expensive, requires a camera monitoring the lips that is fixed to a view of the speaker's head, and it 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 in terms of low reliability. In the following sections, the approach is reported of recording activity of the facial muscles (fEMG) for determining silently commands from a human speaker.

Earlier work reported by the authors have demonstrated the use of multi-channel surface electromyogram (SEMG) to identify the unspoken vowel based on the normalized integral values of facial EMG during the utterance, and this construction had been tested with native Australian English speakers. The main concern with such systems is the difficulty to work across people of different backgrounds, and the main challenge is the ability of such a system to work for people of different languages – native ones as well as foreign ones. Consequently, in this particular work the error in classification of the unvoiced English and German vowels by a group of German native speakers are compared. Hence, this investigation covers the application case of two different languages used by native speakers, and the case of speakers talking and commanding in a foreign language.

2 THEORY

This research aims to recognize the multi-channel surface electromyogram of the facial muscle with speech and identify the variation in the accuracy of classification for two different languages, German and English. Articulatory phonetics considers the anatomical detail of the utterance of sounds. This requires the description of speech sounds in terms of the position of the vocal organs, and it is convenient to divide the speech sounds into vowels and consonants. The consonants are relatively easy to define in terms of 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 (Parsons, 1986). When considering speech articulation, the shapes of the mouth during speaking vowels remain constant while during consonants the shape of the mouth changes.

2.1 Face Movement and Muscles Related to Speech

The human 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 (Lapatki et al., 2003). When using facial SEMG to determine the shape of lips and mouth, there is the issue of the proper

choice of muscles and the corresponding location of the electrodes, and also the difficulty of cross talk due to the overlap between the different muscles. (Chan et al., 2002) demonstrated the presence of speech information in facial myoelectric signals using an SEMG based system. (Kumar et al., 2004) have demonstrated the use of SEMG to identify unspoken sounds under controlled conditions. More algorithmic details for dedicatedly classifying facial muscle activity during vowel-based speech previously were reported in (Arjunan et al., 2006).

Applying 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.

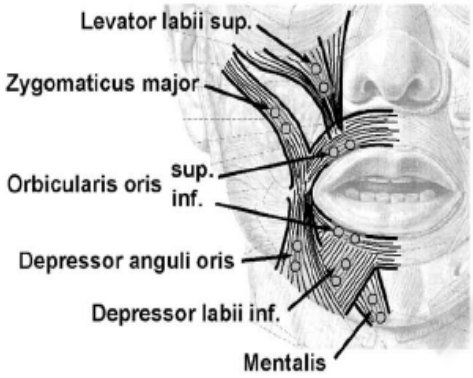


Figure 1: Topographical location of facial muscles [Source: (Lapatki et al., 2003)].

In this study, only the following four facial muscles have been selected: *Zygomaticus Major*, *Depressor anguli oris*, *Masseter* and *Mentalis* (Fridlund and Cacioppo, 1986). The placement of electrodes and location of these muscles are shown in Figure 1. With the variation in speed and pronunciation of speaking, and the length of each sound in different languages, it is difficult to determine an appropriate window in time domain for best signal analysis. When the properties of the signal are time varying, identifying suitable features for classification will be less robust.

2.2 Features of SEMG

Surface electromyogram (SEMG) is a gross indicator of the muscle activity and is used to identify force of muscle contraction, associated movement and posture. SEMG is a complex non-stationary signal. The strength of SEMG is a good measure of the strength of contraction of the muscle, and it can be related to the movement and posture of the corresponding part of the body (Basmajian and Deluca, 1985). 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, and thus the strength of the force of muscle contraction. The issue regarding the use of SEMG to identify speech is the large variability of SEMG activity pattern associated with a phoneme of speech (Basmajian and Deluca, 1985). 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 the issue of variation in speed and pronunciation of vowels, 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 normalised values of the integral of RMS of SEMG from the different muscles to reduce the large inter-experimental variation.

2.3 Statistical Analysis using Cross-validation

Cross-validation is the statistical practice of partitioning a sample of data into subsets such that the analysis is initially performed on a single subset, while the other subset(s) are retained for subsequent use in confirming and validating the initial analysis.

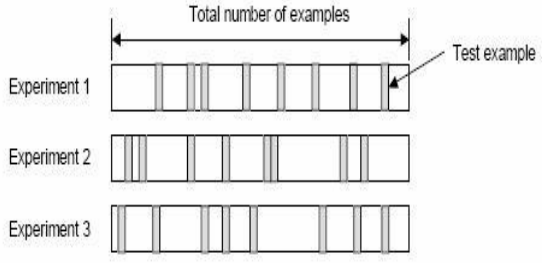


Figure 2: Random sub-sampling cross validation. [Source: (Gutierrez-Osuna, 2001)].

The initial subset of data is called the *training set*; the other subset(s) are called *validation or testing sets*. The *holdout method* is the simplest kind of cross validation. The data set is separated into two sets, called the training set and the testing set. The function approximator fits a function using the training set only. Then the function approximator is

asked to predict the output values for the data in the testing set (it has never seen these output values before). The obtained errors are accumulated as before to determine the mean absolute test set error, which is used to evaluate the model. For having a random selection of training and testing data sets, *Random Sub sampling Cross-validation* method was used. Random Sub sampling performs *k* data splits of the dataset as shown in Figure 2. Each split randomly selects a (fixed) number of examples without replacement. For each data split we retrain the classifier from scratch with the training examples and estimate with the test examples. The true classification accuracy is obtained as the average of the separate estimates (Gutierrez- Osuna, 2001). The training and testing was done using Artificial Neural Network architecture with back propagation algorithm.

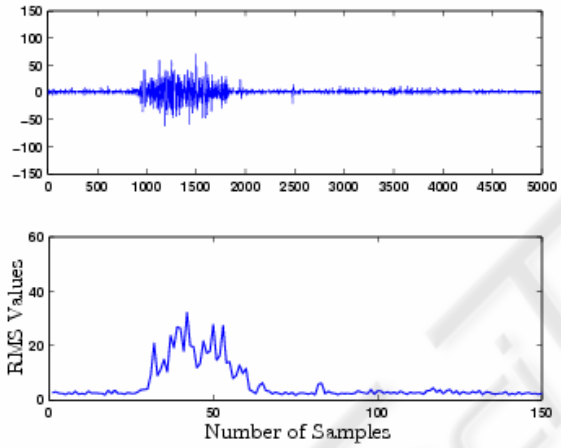


Figure 3: Recorded raw EMG signal and its RMS plot.

3 METHODOLOGY

Experiments were conducted to evaluate the performance of the proposed speech recognition from facial SEMG for two different languages, German and English. The experiments were approved by the Human Experiments Ethics Committee of the University. In controlled experiments, participants were asked to speak while their SEMGs were recorded. The SEMG recordings were visually observed, and all recordings with any artefacts – typically due to loose electrodes or movement – were discarded. During these recordings, the participants spoke three selected English vowels (/a/,/e/,/u/) and three selected German vowels (/a/,/i/,/u/). Each vowel was spoken separately such that there was a clear start and end of its utterance. The experiment was repeated ten times for each language. A suitable resting time was granted to the speakers between each experiment. The participants were asked to vary their speaking speed and style to obtain a wide training set.

3.1 EMG Recording and Processing

In an earlier reference investigation, three male volunteers participated who are English native speakers, while in the present investigation, one female plus two male volunteers participated in the experiments. All the participants in this second experimental run were native speakers of German with English as their second language.

Four channel facial SEMG was recorded using the recommended recording guidelines (Fridlund and Cacioppo, 1986). A four channel, portable, continuous recording MEGAWIN instrument (Mega Electronics, Finland) was used for this purpose. Raw signal was recorded at a rate of 2000 samples per second. Ag/AgCl electrodes (AMBU Blue sensors from MEDICOTEST, Denmark) were mounted on appropriate locations close to the selected facial muscles. The recordings were visually observed, and all recordings with any artefacts were discarded. Figure 3 shows the raw EMG signal recording, and its RMS values plotted as a function of time domain denoted by the corresponding sample number.

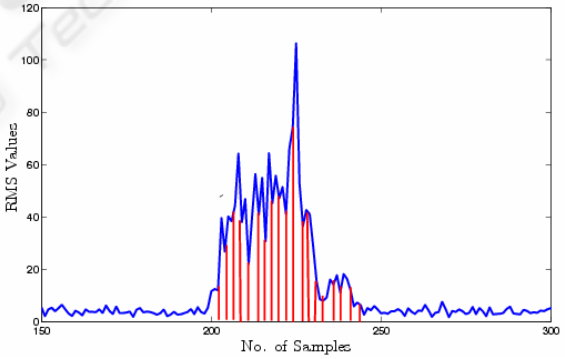
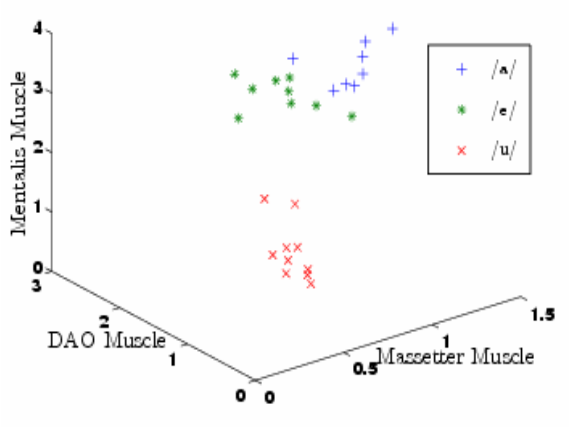


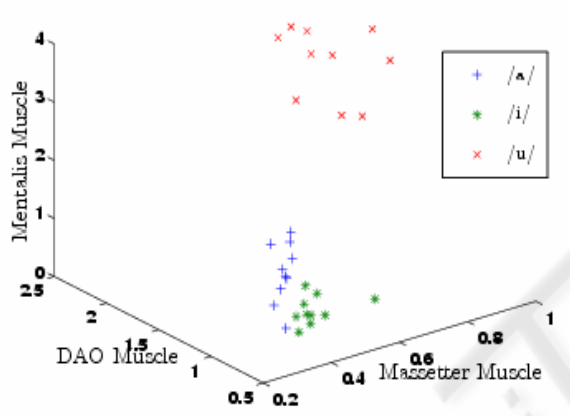
Figure 4: Example RMS integration of SEMG.

3.2 Data Analysis

The first step in the analysis of the data was to identify the temporal location of the muscle activity. Moving root mean square (MRMS) of the recorded signal with a threshold of 1 sigma of the signal was applied for windowing and identifying the start and the end of the active period (Freedman et al., 1997). A Window size of 20 samples corresponding to 10 ms was used for computing the MRMS.



(a)



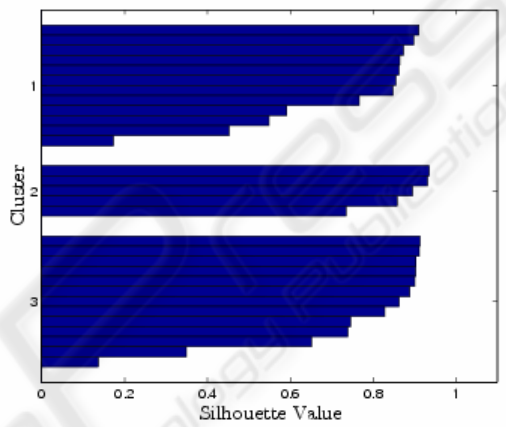
(b)

Figure 5: 3-D plot of the normalised IRMS values for (a) English Vowels, and for (b) German Vowels.

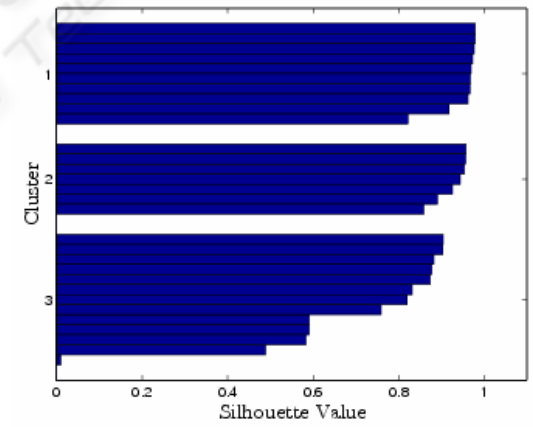
The start and the end of the muscle activity were also confirmed visually. The next step was to parameterise the SEMG for classification of the data. MRMS values of SEMG between the start and the end of the muscle activity was integrated for each of the channels. This provided a four long vector corresponding to the overall activity of the four channels for each vowel utterance. This data was normalised by computing a ratio of integrated MRMS of each channel with respect to channel number one. This ratio is indicative for the relative strength of contraction of the different muscles and reduces the impact of inter-experimental variations. The outcome of this step was a vector of length three corresponding to each utterance. Figure 4 is an example of the computation of the integral of RMS of SEMG. For computing the integral of RMS of SEMG, Durand's rule (Beyer, 1987) was used, because it produces approximations that are more accurate, since these represent a straightforward family of numerical integration techniques.

3.3 Classifying of Normalised Features of Facial SEMG

For classification, parameterised SEMG data resulting in a vector with three measures for each utterance was used. The first step was to determine, if this data is separable.



(a)



(b)

Figure 6: Silhouette plot of the normalised IRMS values for (a) English Vowels, and (b) German Vowels.

After confirming this, the next step was to determine whether the data is *linearly* separable. A supervised neural network approach was used for the separation step. The advantage of using a neural network is that such networks can be applied

without the assumption for a possible linear separation of the data. For this purpose, the data from the ten experiments for each vowel uttered by one individual participant was divided into two groups - training and test data.

This was repeated for English and German language separately. The ANN consisted of two hidden layers with 20 nodes in both layers. Sigmoid function was used as the threshold decision. ANN was trained with gradient descent algorithm using a momentum with a learning rate of 0.05 to reduce the likelihood of local minima. Finally, the trained ANN was used to classify the test data. This entire process was repeated for each of the participants. The performance of these integral RMS values was evaluated by comparing the accuracy in the classification during testing. The accuracy was calculated as ratio of the percentage of correctly classified data points and the total number of data points in the class.

The next step in the classification of this data was to test, whether the data was linearly separable. Taking advantage of the three dimension in the data, three axis plot was produced. In this, data points representing each vowel were given a specific colour and distinct symbol for visual inspection. Figure 5 shows example of such a plot, for each of the investigated languages. The K-means clustering technique was performed to test the data for linear separability. To get an idea of how well-separated the resulting clusters are, a silhouette plot was made using the cluster indices output from k-means. The silhouette plot in Figure 6 displays a measure of closeness of each point in one cluster to points in the neighbouring clusters.

3.4 Statistical Analysis of Classification Accuracy

Random Sub sampling cross-validation method was used to determine the mean classification accuracy of the normalised features of facial SEMG. The training and testing of different random sub samples using ANN was repeated for different times. The final classification accuracy is the average of individual estimates as in Eqn.1.

$$A = 1/K \times \sum_{i=1}^{K} (c/n) \qquad (1)$$

where c = number of correctly classified utterances
n = total number of utterances
K = total repetition count of training and testing

4 RESULTS AND OBSERVATIONS

The linear separation of normalised IRMS values of different vowels was tested using three dimensional plot and silhouette plot. It is observable from the 3-D plots in Figure 5, that there appears distinct clustering of the data based on the vowel uttered for *both* languages.

Table 1: Mean classification accuracy for English vowels.

Vowels	Mean Classification accuracy		
	Participant 1	Participant 2	Participant 3
/a/	73.3%	83.3%	80.0%
/e/	76.7%	76.7%	83.3%
/u/	100.0%	100.0%	100.0%

Table 2: Mean classification accuracy for German vowels.

Vowels	Mean Classification accuracy		
	Participant 1	Participant 2	Participant 3
/a/	86.7%	83.3%	83.3%
/i/	96.7%	80.0%	76.7%
/u/	100.0%	100.0%	100.0%

This is also verified using k-means Silhouette plot (Figure 6): it is clear that most points have a large silhouette value, indicating that the clusters are separated from each other and this suggests that there exists a linear separation of the data. The average silhouette values for English vowels and German vowels are 0.7634 and 0.8441 respectively. This shows that the linear separation of data is stronger in German vowels (native language of the speaker) than English vowels (foreign language). Table 1 shows the ANN classification results on the test data using weight matrix generated during training for English vowels, and Table 2 lists these values for German vowels. These results indicate that the mean classification accuracy of the integral RMS values of the EMG signal yields better recognition rate of vowels for 3 different participants, when it is trained individually. The results indicate that this technique can be used for the classification of vowels for the native and foreign language – in this case – English and German. This suggests that the system is able to identify the differences between the styles of speaking of different people at different times for different languages.

4.1 Variation in Classification Error for Native and Foreign Language

Figure 7 shows the variation in error rate for German and English vowels. The error rate in classification accuracy for a foreign language (English) is marginally high when compared with the native language (German). This is due to the muscle pattern remaining same during the utterance of the native

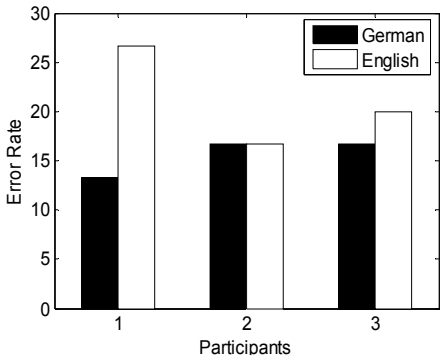


Figure 7: Error bar plots for classification of English and German Vowel - /a/.

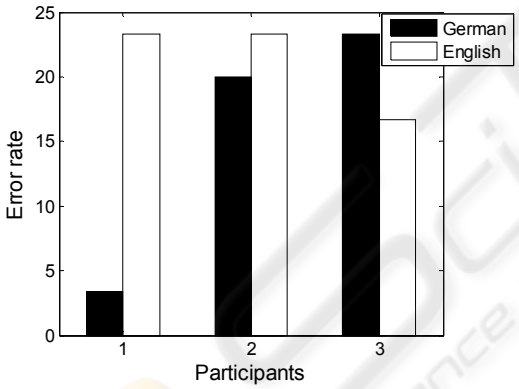


Figure 8: Error bar plots for classification rates of the German vowel - /i/ and the English Vowel - /e/.

language and changes during the utterance of the foreign language. The variation is high for German vowels /a/, /i/ and English vowels /a/, /e/ and there is no variation for the vowel /u/ in both German and English language. This can also be seen from the data pattern for both the languages (3-D Plot) in Figure 5.

5 DISCUSSION

The results indicate that the proposed method using activities of facial muscles for identifying silently spoken vowels is technically feasible from the view point of error in identification. The investigation reveals the suitability of the system for English and German, and this suggests that the system is feasible when used for people speaking their own native language as well as a foreign language.

The results also indicate that the system is not disturbed by the variation in the speed of utterance. 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). Up to now, this method has only been tested for limited vowels. Vowels were considered at first, 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 and reliable method for classifying vowels of single user without regard to speaking speed and style in different times for different languages. It should be pointed out that this method at this stage is not being designed to provide the flexibility of regular conversation language, but for a limited dictionary only, which is appropriate for simple voice control systems. 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.

6 CONCLUSIONS

This work describes a silent vowel based speech recognition approach that works with measuring the facial muscle contraction using non-invasive SEMG. Application of this includes, e.g., removal of any disambiguity caused by the acoustic noise for human computer interface or computer based speech analysis. The presented investigation focused on classifying English and German vowels, because pronunciation of vowels results in stationary muscle contraction as compared to consonants.

The system has been tested with a very small set of phonemes, where the system has been successful. The recognition accuracy is high, when it is trained and tested for a dedicate user. It should be pointed out that this method at this stage is not yet designed to provide the flexibility of regular conversation language, but for a limited dictionary only, which is appropriate for simple voice control systems. This

method has to be enhanced for large set of data with many subjects in future.

One basic application for such a system is for disabled user to give simple commands to a machine, which would be a helpful and typical application of HCI. Future applications of such a system, e.g., cover Internet access for people on the move while using their mobile devices, or when they are in public places. Further possibilities especially include applications for telephony, defence problems, and improvement of speech-based computer control in general in any noisy environment.

REFERENCES

Arjunan, S.P., Kumar, D.K., Yau, W.C., Weghorn, H. (2006). Unvoiced speech control based on vowels detected by facial surface electromyogram. In *Proceedings of IADIS international conference e-Society 2006*, Dublin, Ireland, Vol. I, pp. 381-388.

Basmajian, J. V. and Deluca, C. J. (1985). *Muscles Alive: Their Functions Revealed by Electromyography*. Williams and Wilkins, Baltimore, fifth edition.

Beyer, W. H. (Ed.) (1987). *CRC Stndard Mathematical Tables*. CRC press, Boca Raton, 28th edition. p. 127

Chan, A., Englehart, K., Hudgins, B., and Lovely, D. (2002). A multi-expert speech recognition system using acoustic and myoelectric signals. In *Proceedings of 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society*, Ottawa, Canada, Vol. 1, pp. 72-73. IEEE.

Freedman, D., Pisani, R., and Purves, R. (1997). *Statistics*. Norton College Books, New York, third edition.

Fridlund, A. and Cacioppo, J. (1986). Guidelines for human electromyographic research. In *Journal of Psychophysiology*, Vol. 23, No. 4, pp. 567-589. The Society for Psychophysiologial Research.

Gutierrez-Osuna, R. (2001). Lecture 13: Validation. In *http://research.cs.tamu.edu/prism/lectures/iss Last access: October 2006*. Wright State University.

Kumar, S., Kumar, D., Alemu, M., and Burry, M. (2004). EMG based voice recognition. In *Proceedings of Intelligent Sensors, Sensor Networks and Information Processing Conference*, Melbourne, Australia. IEEE.

Lapatki, B. G., Stegeman, D. F., and Jonas, I. E. (2003). A surface EMG electrode for the simultaneous observation of multiple facial muscles. In *Journal of Neuroscience Methods*, Vol. 123, No. 2, pp. 117-128.

Manabe, H., Hiraiwa, A., and Sugimura, T. (2003). Unvoiced speech recognition using SEMG- mime speech recognition. In *ACM Conference on Human Factors in Computing Systems*, Ft.Lauderdaler, Florida, USA, pp. 794-795.

Parsons, T. W. (1986). *Voice and speech processing*. McGraw-Hill Book Company, New York, first edition. Ursula, H. and Pierre, P. (1998). Facial reactions to emotional facial expressions: affect or cognition? In *Cognition and Emotion*, Vol. 12 No.4.