

Vowel recognition of English and German language using Facial movement(SEMG) for Speech control based HCI

Sridhar P Arjunan¹

Hans Weghorn²

Dinesh K Kumar¹

Wai C Yau¹

¹School of Electrical and Computer Engineering, RMIT University, Melbourne, Australia

²Information Technology, BA-University of Cooperative Education, Stuttgart, Germany

Email: sridhar_arjunan@ieee.org, weghorn@ba-stuttgart.de, dinesh@rmit.edu.au,

Abstract

This paper examines the use of facial muscle activity (Surface Electromyogram) to recognise speech based commands in English and German language 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. 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 foreign language. To overcome this difficulty, the paper reports measuring the relative activity of the articulatory muscles for recognition of unvoiced vowels of English and German languages. In these investigations, 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 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 gives high recognition rate when used for each of the participants for both of the languages. The investigations also show that the system is easy to train for a new user. The visual inspection of the plot of the experimental data suggests the formation of clusters. The results suggest that such a system is reliable for simple vowel based commands for human computer interface when it is trained for the user, who can speak one or more languages and for the people who have speech disability.

Keywords: Surface Electromyogram, Speech control, HCI, ANN.

1 Introduction

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 fo-

cused on three human factors of body functions: speech, bioelectrical activity and facial expressions. 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 1998)

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.

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 (Manabe 2003) or by relating the movement and shape to the muscle activity (Chan 2002, Kumar 2004). 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.

Earlier work reported by the authors have demonstrated the use of multi-channel surface electromyo-

gram (SEMG) to identify the unspoken vowel based on the normalized integral values of SEMG during the utterance. The main common 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 native languages. Earlier work by the authors had tested the system for native Australian English speakers. This paper compares the error in classification of the unspoken English and German vowels by a group of German native speakers.

2 THEORY

The purpose of this research is to classify the surface recordings of the facial muscle activity with speech. For this analysis, 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. It is important to identify the anatomical details of speech production for analysing the shape of the mouth and the muscle activity with speech.

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 (Thomas 1986). 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 orofacial functions such as speech, mastication, swallowing and mediation of motion (Lapatki 2003). 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 (Thomas 1986).

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 (Basmajian 1985). Using an SEMG based system, Chan et al (Chan 2002) demonstrated that the presence of speech information in facial myoelectric signals. Kumar et al (Kumar 2004) 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 (Veldhuizen 2003) 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 (Fridlund 1986). The location of these muscles are shown in Figure 1.

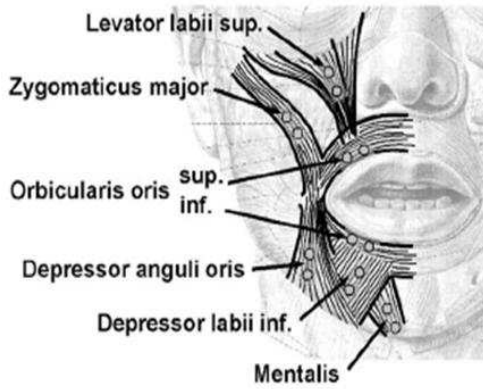


Figure 1: Topographical location of facial muscles [Source: (Lapatki 2003)]

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 (Chan 2002). 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 (Basmajian 1985).

The vowels correspond to stationary muscle activity, the muscle activity pre and post the vowel is non-stationary. 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 is overcome.

3 Methodology

Experiments were conducted to evaluate the performance of the proposed speech recognition from facial EMG for different languages, German and English. The experiments were approved by the Human Experiments Ethics Committee of the University. Controlled experiments were conducted where the par-

Vowels					
Short	a	e	i	o	u
	[a]	[ɛ]	[ɪ]	[ɔ]	[ʊ]
Long	a/aa/ah	e/ee/eh	i/ih/ie	o/oo/oh	u/uh
	[a:]	[e:]	[i:]	[o:]	[u:]

Figure 2: Pronunciation of German vowels

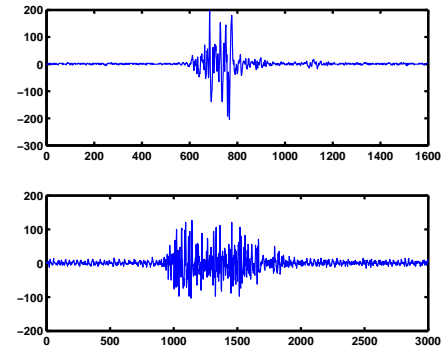


Figure 3: Raw SEMG signal recorded during different experiments

ticipant was asked to speak while their SEMG was recorded. The SEMG recordings were visually observed, and all recordings with any artifacts - typically due to loose electrodes or movement- were discarded. During these recordings, the subjects 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 given between each experiment. The participants were asked to vary their speaking speed and style to obtain a wide based training set. The pronunciation of German vowels (Ager 2006) is shown in Figure 2.

3.1 EMG Recording and Processing

In previous investigation, three male volunteers speaking English participated and in the present investigation, two male and one female volunteers participated in the experiments. All the participants in this experiment were native speakers of German with English as their second language. Four-channel facial SEMG was recorded using the recommended recording guidelines (Fridlund 1986).

A four channel, portable, continuous recording MEGAWIN instrument (from Mega Electronics, Finland) was used for this purpose. Raw signal was recorded at a rate of 2000 samples/second. Ag/AgCl electrodes (AMBU Blue sensors from MEDICOTEST, Denmark) were mounted on appropriate locations close to the selected facial muscles, which 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 experiments. The recordings were visually observed, and all recordings with any artifacts were discarded. Figure 3 shows the raw SEMG signal recorded during different experiments by changing the speed and style of utterance, plotted as a function of time (sample number).

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 with a threshold of 1 sigma of the

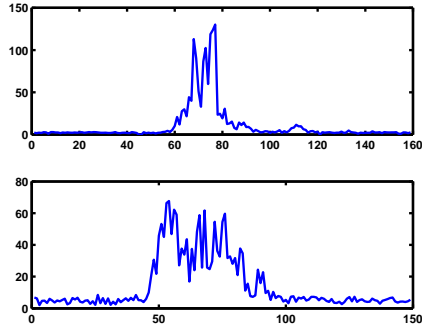


Figure 4: RMS plot of the recorded EMG signals

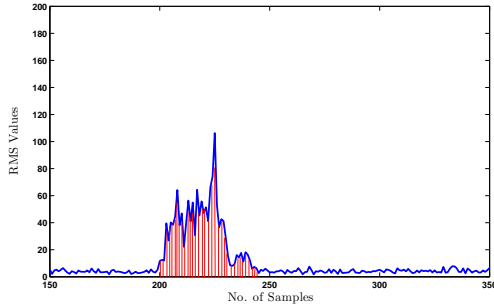


Figure 5: An example of the computation of the integral of RMS of SEMG

signal was applied for windowing and identifying the start and the end of the active period (David 1997). Window size of 20 samples corresponds to 10 msec and was used as the size of the window for computing the MRMS. The start and the end of the muscle activity were also confirmed visually. Figure 4 shows the plot of RMS values of the different recorded EMG signals. The next step is to parameterise the SEMG for classification of the data. To overcome the difference between the speed of utterance during different experiments, and difference between different experiments in the absolute magnitude of the recordings, the data was integrated and normalised. MRMS of the envelope 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 with respect to channel 1 by computing a ratio of integrated MRMS of each channel with respect to channel 1. This ratio is indicative for the relative strength of contraction of the different muscles and reduces the impact of inter-experiment variations. The outcome of this step was a vector of length three corresponding to each utterance. Figure 5 is an example of the computation of the integral of RMS of SEMG.

For computing the integral of RMS of SEMG, Durand's rule (Eric 2006) was used, because it produces approximations that are more accurate and a straightforward family of numerical integration techniques. A simplified block diagram of methodology shown in Figure 6, explains the process of the analysis.

3.2 Classification of SEMG Data

Parameterization of SEMG data results in a vector with three measures for each utterance. The first step in classification of data was to determine if this data was separable. After confirming this, the next step

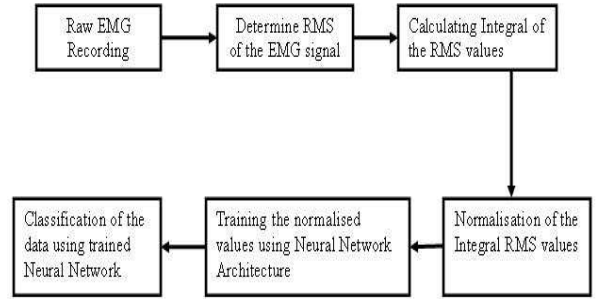


Figure 6: A simplified block diagram of methodology

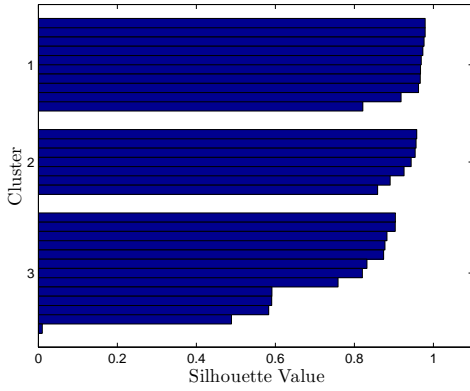
undertaken was to determine whether the data is linearly separable. To determine whether the data is separable, supervised neural network approach was used. The advantage of using such a neural network is that neural networks can be applied without the assumption for linear separation of the data. For this purpose, the data from the ten experiments for each participant was divided into two equal groups - training and test data. This was repeated for English and German language separately. An over-sized neural network was used to ensure identifying the separation of the data.

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 momentum with a learning rate of 0.05 to reduce 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 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.

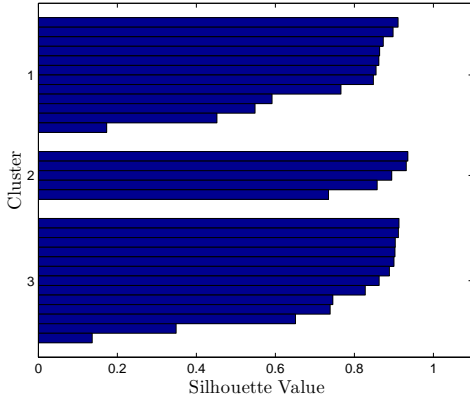
The next step in the classification of this data was to test, whether the data was linearly separable. Taking advantage of the three dimension in the data, three axis plot was produced. In this, data points representing each vowel were given a specific colour and distinct symbol for visual inspection. Figure 8 and Figure 9 show examples 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 7 displays a measure of closeness of each point in one cluster to points in the neighbouring clusters. Unsupervised clustering does not demonstrate linear separability of the data with no temporal information and with a prior knowledge of the targets against the inputs. Supervised back-propagation neural network is the most convenient to use as a classifier, the authors are aware that such a classifier may in some cases be sub-optimum. The advantage of ANN approach is that ANN is easy to be trained by a user to configure the system for the individual.

4 RESULTS AND OBSERVATIONS

The linear separation of normalised integral RMS values of different vowels was tested using three dimensional plot and silhouette plot. It is observable from the 3-D plots in Figure 8 and Figure 9, that there appears distinct clustering of the data based on the vowel uttered for *both* languages. This is also verified using k-means Silhouette plot (Figure 7), it is clear



(a)



(b)

Figure 7: Silhouette plot of the normalised IRMS values (a) English Vowels (b) German Vowels

Table 1: Classification results for different participants uttering English vowels

Vowel	Correctly Classified Vowels		
	Participant 1	Participant 2	Participant 3
/a/	3(60%)	4(80%)	4(80%)
/e/	4(80%)	4(80%)	4(80%)
/u/	5(100%)	5(100%)	5(100%)

that most points have a large silhouette value, indicating that the clusters are separated from each other and it suggests that there exists 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).

The results of testing the ANN on the test data using weight matrix generated during training are tabulated in Table 1 for English vowels and Table 2 for German vowels. These results indicate an overall average accuracy of 86%, 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.

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.

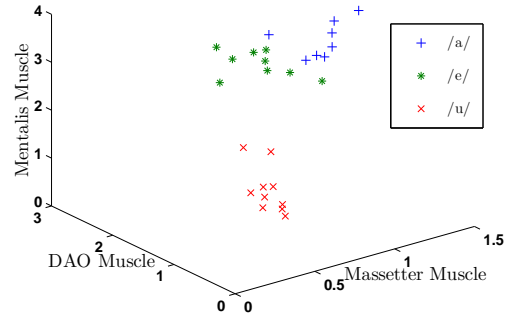


Figure 8: 3-D plot of the normalised IRMS values of English vowels

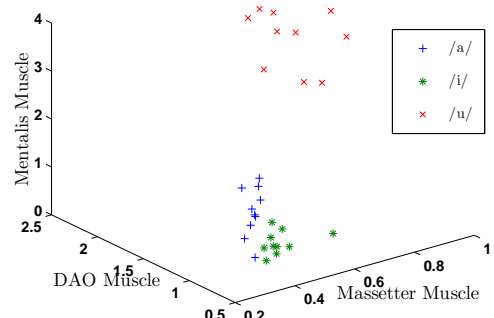


Figure 9: 3-D plot of the normalised IRMS values of German vowels

5 DISCUSSION

The results indicate that the proposed method that uses 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). 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.

Table 2: Classification results for different participants uttering German vowels

Vowel	Correctly Classified Vowels		
	Participant 1	Participant 2	Participant 3
/a/	4(80%)	4(80%)	4(80%)
/i/	5(100%)	4(80%)	4(80%)
/u/	5(100%)	5(100%)	5(100%)

els of single user without regard to the speaking speed and style in different times for different languages. 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, 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 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 and is suitable for two languages - English and German. Application of this include e.g. removal of any disambiguity caused by the acoustic noise for human computer interface or computer based speech control and 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 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. The system has been tested with a very small set of phonemes, where the system has been successful. A broad variety of applications could benefit from this technology: 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, defence problems and improvement of speech-based computer control in noisy environments.

7 Acknowledgments

We would like to thank for the financial support of the Landesstiftung Baden-Württemberg GmbH in Stuttgart, Germany, who co-financed the visiting research placement of the main author of this paper. The supplied funding by the Baden-Württemberg Stipendium enabled that this particular investigation on multi-language voice control could be performed in an efficient manner.

References

- Basmajian, J.V. & DeLuca, C.J. (1985), *Muscles Alive; Their Functions Revealed by Electromyography*, Fifth edition, Williams & Wilkins, Baltimore.
- Thomas W. Parsons. (1986), *Voice and speech processing*, McGraw-Hill, First edition, New York.
- David Freedman, Robert Pisani & Roger Purves. (1997), *Statistics*, Norton College Books, Third Edition, New York.
- Chan, D.C., Englehart, K., Hudgins, B. & Lovely, D. F. (2002), 'A multi-expert speech recognition system using acoustic and myoelectric signals', in *Proceedings 24th Annual Conference and the Annual Fall Meeting of the [Biomedical Engineering Society] EMBS/BMES Conference*, Ottawa, Canada, Vol. 1, pp. 72-73.

- Kumar, S., Kumar, D.K., Alemu, M. & Burry, M. (2004), 'EMG based voice recognition', in *Proceedings of Sensor Networks and Information Processing*, Melbourne, Australia.
- Manabe, H., Hiraiwa, A. & Sugimura, T. (2003), 'Unvoiced speech recognition using SEMG - Mime Speech Recognition', *ACM Conference on Human Factors in Computing Systems*, Ft.Lauderdale, Florida, USA, pp. 794-795.
- Veldhuizen, I.J.T., Gaillard, A.W.K. & de Vries, J. (2003), 'The influence of mental fatigue on facial EMG activity during a simulated workday', *In Journal of Biological Psychology*, Vol. 63, No. 1, pp. 59-78.
- Fridlund, A.J., & Cacioppo, J.T. (1986), 'Guidelines for Human Electromyographic research', *In Journal of Psychophysiology*, Vol. 23, No. 4, pp. 567-589.
- Lapatki, G., Stegeman, D. F. & 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.
- Ursula, H., Pierre, P. & Sylvie, B. (1998), 'Facial Reactions to Emotional Facial Expressions: Affect or Cognition?', *Cognition and Emotion*, Vol. 12, No. 4.
- Eric W. Weisstein (2006), Durand's Rule, From MathWorld- A Wolfram Web Resource, <http://mathworld.wolfram.com/DurandsRule.html>. Accessed August 2006.
- Ager, S. (2006), Information about German language, <http://www.omniglot.com/writing/german.htm>. Accessed August 2006.