Electromyography Based Word Recognition for Silent Speech Interface

Geeta N. Sonawane[1] Mrs. A.N.Shewale[2] Gunjan G. Gujarathi[3]
[1]Research Student [2]Associate Professor [3]Assistant Professor
[1,2,3]SGDCOE, NMU, Maharashtra, India

Abstract— One of the electronic systems that enable to communicate by speech without an audible acoustic signal is Silent Speech Interface. The work investigates the usability of muscle contractions detected by surface electromyography (EMG) sensors as an input channel for Silent Speech Interface. Therefore, the technology enables speech recognition to be applied to silently mouthed speech. The work describes the outcomes in using artificial neural network to recognize and classify human speech based on EMG signals which are captured at the facial muscles, records the activity of the human articulatory apparatus and thus allows tracing back a speech signal even if it is spoken silently. Since speech is captured before it gets airborne, the resulting signal is not masked by ambient noise. The preliminary results demonstrate that the proposed technique yields high recognition rate for classification of unvoiced words using SEMG features. The output of Silent Speech Interface has the potential to overcome major limitations of conventional speech driven interfaces: it is not prone to any environmental noise, allows to silently transmitting the confidential information, and does not disturb bystanders. The results demonstrate that the system is easy to train for a new user. This work forms the basis for further researches to use EMG signals to improve large training dataset and uses of different languages based model.

Key words: Electromyography (EMG), SEMG

I. INTRODUCTION

Human speech communication usually takes place in presence of complex acoustic backgrounds with competing voice, environmental sound sources and ambient noise. In such condition it is quite difficult to for the human speech to remain robust. One of the electronic systems that enable to communicate by speech without an audible acoustic signal is Silent Speech Interface [1]. In contravention of success, speech-based technologies still face the challenges like recognition performance degrades significantly in the presence of noise and confidential or private communication in public places is jeopardized by audible speech. Both of these challenges are addressed by SSI.There are several electromyographic approaches used in which acoustic speech recognition is replaced by silent speech recognition [2].Electromyography is thestudy of muscle functionbased on the examinationand analysis of theelectrical signals thatemanate from the muscles [3].These signals are formed by physiological variations in the state of muscle fiber membranes. Myoelectric signals are measurable signals which appear during muscle activation. During muscle contraction, small electrical currents are generated by the exchange of ions across muscle fiber membranes [4].

Using audible speech signal a confidential conversation with or through a device becomes impossible. In libraries or during meetings talking can be extremely disturbing to others. Performance is also degrades when sound production limitations occur, like under water. At last for speech handicapped people, that is those without vocal cords the conventional speech-driven interfaces cannot be used. In the future, SSIs may overcome these limitations by allowing people to generate natural sounding speech from the movements of their tongue and lips. The example of EMG is shown in Fig. 1.

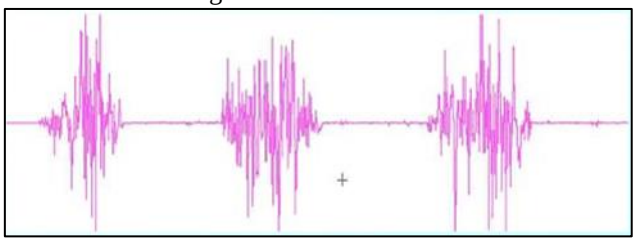

Fig. 1: Electromyography

Bio-medical system is the fast growing technology proposed the use of electromyography signal in which the acoustic speech recognition is substituted by silent-speech recognition. EMG signals are detects from the surface and gives the time domain features associated with it. Significant motions are extracted using suitable technique. The obtained signals are non-liner, non-stationary, complexity and have large variation, which creates difficulty in analyzing EMG signal.The application areas of EMG based SSI are confidential, robust, non-disturbing speech recognition for human machine interface and transmission of articulatory parameters like a mobile telephone for example silently speaking text messages.

This paper organizes as in section II literature review on facial and visual modalities used for speech recognition addressed.In section III development of proposed system for word from articulatory muscles movement without using acoustic information. In section IV experimental results are determined for different word data. Finally conclusions and some direction for further research will be given in section V.

II. LITERATURE SURVEY

Generally for non-acoustic speech recognition two types are used namely visual and facial muscles activity based on SEMG for identification of silent speech. Facial approach can be classified on the basis of sampled data like vowel, syllable, digits, and words, sentences while visual approach intends visualization of acoustic and articulatory features.

A. Facial Approach

The human face can communicate a variety of information including subjective emotion, commutative intent, and cognitive appraisal. The facial musculature is a three dimensional assembly of small, pseudo-independently controlled muscular lips performing a variety of complex facial functions such as speech, mastication, swallowing and mediation of motion. 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. Many researches had

done which uses facial modality and recognized vowel, digits, words, syllable or sentence.

English vowels are building block in modern speech. The author Sanjay Kumar et.al employed the EMG vowels data extracted from three articulatory facial muscles using neural networks [5].Their work had reports the use of SEMG with success to identify the sub-auditory sounds using neural networks. For EMG recording and processing three male subjects and the AMLAB work station was used. The subjects' spoken five English vowels for three times were recorded and observed using three facial EMG simultaneously. Signal Processing of SEMG intend the root mean square of the signals indicates the power generated by the muscles. Back Propagation type Artificial Neural Network used for speech recognition from EMG to overcome the drawback of the standard ANN architecture. The three RMS EMG values were the inputs to the ANN while the output of the ANN was one of the five vowels. They described promising result with the system could classifies the five vowels with an accuracy. While the study required bigger experimental population.

Usually syllables composed by a consonant followed by a vowel [2], divided into five groups as in Table 1.

Vowels	a	e	i	o	u
Labials	pa	pe	pi	po	pu
Dentals	ta	te	ti	to	tu
Palatals	ya	ye	yi	yo	yu
Velars	ka	ke	ki	ko	ku
Alveolars	la	le	li	lo	lu

Table 1: Complete Set of Syllables

Author focused on syllable of Spanish language based on EMG signal recorded in facial muscles [2]. They proposed method to obtain a natural speech recognizer for the recognition of the syllables. Syllables are simply voice hits and correspond to abrupt muscle movements. For experimental purpose, the EMG signals corresponding to 50 examples of each of the Spanish syllables recorded. They used the boosting algorithm AdaBoost as classifier. Using the software Weka the training and classification processes was carried out. The result suggested a high performance and potential of the recognition system given the large number of classes involved in the problem. The feature vector whose components represented different global characteristics extracted from each articulated syllable signal. Using feature vectors as input the classifier based on boosting was trained. Further to build the complete Spanish words required.

Jun Wang discussed the prime requirement of word recognition for SSI [7]. The paper shows the application of articulation-based SSI, can be used to produce synthetic speech to enable voiceless patients by using their lips and tongue and also used in command-and-control systems. For that they developed word recognition algorithm. The designed articulation-based SSI contains three major components: data acquisition, online word recognition, and sound playback or synthesis. EMG signal used for data acquisition detect the motion of sensors applied on a speaker's lips and tongue. This paper focused on online word recognition. The segmentation and identification were operated together in a variable-size sliding window for whole-word recognition algorithm. Symbolic representation technique has been widely used in time-series data pattern analysis. In this study, to discretize the lips and tongue motion time series data SAX was used. The word recognition algorithm's performance was evaluated by accuracy recognition and time processing. The result was satisfactory still determination of the optimal parameters like candidate thresholds automatically for online recognition is required.

To recognize the word, one more method described in [10]. The study states that by using only one single pair of surface electrodes 92% accuracy for six acoustic words possible to obtain for measuring and classification. With complex dual quad tree wavelet transforms, noise filtered and feature extraction done for recorded EMG signals from the larynx and sublingual areas below jaw. Feature sets for six sub-vocally pronounced words: stop, go, left, right, alpha, omega were trained. They demonstrated discrete task control words approach for recognition. The approach concentrated on the fact that vocal speech muscle control signals must be highly repeatable to be understood by others. MATLAB scripts were developed for signal feature processing. They used a simple mean for representative value. While experimenting, some signals were not recognized by neural net satisfactorily. Result shows that the method was sufficient where discrete word, subject specific, limited control vocabularies applications required. Still generalize trained feature sets to other users, reduce sensitivity to noise and electrode locations, and handle changes in physiological states of the users' remains to work.

To address recognition problems in words, the continuous sentence recognition can becomes the alternate. Authors Michael Wand et al. proposed the first application of the new array technology [8]. They show that the recognition result improved by Independent Component Analysis. The experimental result was slightly worse for without ICA yet repeatability of ICA was not satisfactory. This study introduced multi-channel electrode arrays based an EMG recording system. The test and training sentence were recorded by the English speaker in a quiet room in normal audible speech. The incoming EMG signal channels split into high and low frequency then framing was done. For each channel same process was performed. The data set is of 136 classes with 45 English phonemes at start middle and end parts. The trained acoustic model used for decoding. In comparison with original EMG data the recognition results were better for the ICA-processed signals. Though the methods to distinguish content-bearing signals and noise components yet to be developed. Stan Jou suggested that EMG continuous speech recognition is a system which uses the information from EMG articulatory feature [9]. These articulatory feature classifiers could advantage from the E4 feature that make better to the F-score of the AF classifiers. In experimental setup the recorded speech divided in audible and EMG speech recognizer. In audible Speech Recognizer used Mel-frequency cepstral coefficients with vocal tract length normalization and cepstral mean normalization was used for frame-based feature. Similar to the training of EMG speech recognizer, the AF classifiers were also trained on the EMG signals without speech acoustics. The continuous speech recognition done by stream architecture was a list of parallel feature streams and each of them contains one of the acoustic or articulatory features. Generate the EMG acoustic

model scores for decoding by using information from all streams was combined with a weighting scheme. In the stream architecture test set was divided into two equally-sized subsets in two-fold cross validation. From this inconsistency taken result of further investigation of AF selection is necessary for generalization. In the future, feature selection and weighting schemes were used of the stream architecture.

To improve the real time SSI the author Szu-Chen Jou had taken research for continuous EMG speech recognition system on normal audible speech [10]. This could be use of phoneme based acoustic models and feature extraction methods designed for continuous EMG speech. For audible speech recording they used Broadcast News speech recognizer trained with the Janus Recognition Toolkit. Frame based feature this system used Mel-frequency cepstral coefficients with vocal tract length normalization and cepstral mean normalization is used to get the frame-based feature. All the EMG signals were preprocessed as to estimate the DC offset from the special silence utterances on a per session basis. They model the anticipatory effect by adding frame-based delays to the EMG signals. Also the time-domain mean feature provided additional gain to spectral feature. But even if the spectral features were better and they still very noisy for acoustic model training. The model for channel-specific anticipatory effect which improves the EMG features extraction yet to be designed.

Comparing with former approaches used words sentence as model units, in paper the variations in the EMG signal caused by speaking modes was studied by Matthias Janke et al. [11]. The author suggested this technology the non-acoustic signal was produced and could be used silently. For data acquisition EMG signal were recorded where the position of electrode setting used five channels and captures signals from the levator anguli oris, the zygomaticus major, the platysma, the anterior belly of the digastric and the tongue. Topographical location of facial muscles as shown in Fig. 2;

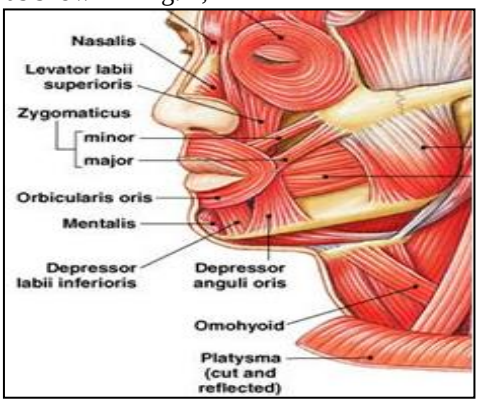

Fig. 2: Topographical Location of Facial Muscles

The feature extractions were based on time-domain features and normalize the frame. They use cross model initialization Cross-Modal Testing that they directly used the base recognizer to decode the silent EMG test set. In Cross-Modal Labeling used trained models from the base recognizer to create a time-alignment for the silent EMG data. Analysis of system computed the ratio of audible EMG and silent EMG PSD of each channel for each frequency bin and took the mean of this ratio over the frequency bins. Also, the calculated WER difference between audible EMG and silent EMG was a measure of EMG recognition performance on audible and silent speech. Experimental purpose it taken relationship between spectral contents of audible and silent speech. With spectral mapping is improved the Cross-Modal Labeling System yields an average WER. In one major study Jun Wang, Ashok Samal, Jordan R. Green introduced Speaker-independent speech recognition method [12]. The across-speaker articulatory normalization based on procures matching for speaker independent silent speech recognition. It had taken the component design of the SSI: real-time articulatory data acquisition, online silent speech recognition and text-to-speech synthesis for speech output as shown Fig. 3;

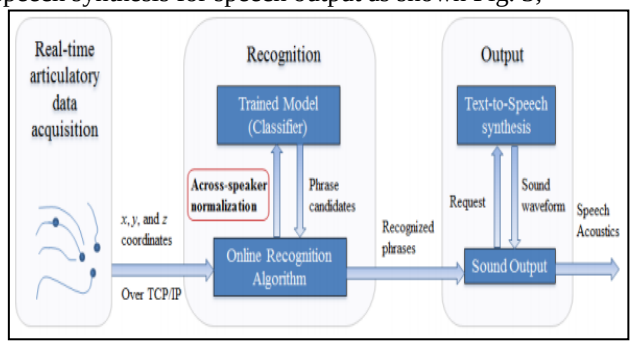

Fig. 3: Conceptual Design of the Speaker Independent Silent Speech Interface

For articulatory normalization of speech recognition they used procures matching, was a robust bi-dimensional shape analysis. In that a shape was represented as a set of ordered landmarks on the surface of an object. In normalization approach transformed each participant's articulatory shape into a normalized shape. It had designed as centroid at the origin, a unit size and aligned to the vertical line that formed the average positions of the upper and lower lips. SVMs used for classification that find separating hyper planes with maximal margins between classes and successfully classifying phonemes, words, and phrases from articulatory movement data. Their experimental results showed consequence of the normalization approach to improving the accuracy of speaker-independent SSI. In future work this approach used in a real-time silent speech interface.

Although the efforts taken on recognition of vowels, words and sentences, digit also essential in daily communication. Till date, the practicability of session dependent speech recognition still limited was suggested by Lena Maier-Hein [13]. The channel dependence of conventional speech recognizer could be a better option for this. The conventional speech recognizer was the result from resulting from the microphone quality, signal transmission of the acoustic signal, and the environmental noise. "Zero" to "nine" ten English digits contained in vocabulary. Total five recording sessions in morning and afternoon on four different days for three subjects were taken. The testing result was considerably worse for across-sessions than within-session testing. The result suggest that methods used in speaker adaptation and conventional speech recognition systems for channel can be used in EMG based speech recognizers for session adaptation. They obtained high average accuracy of word for within-session using seven EMG channels. It was observed that the EMG-based speech

recognition applications on personal devices the speaker independence was not reliable.

B. Visual Modalities

The expressions of speech and emotions play vital role in human interaction, therefore visual and SEMG signals are selected for HCI applications. Hao Li et al.had focused on the visualization of articulators from acoustic signal frame by frame in [14]. To describe human's lips and tongue movements those were more stable method used was Directional relative displacement feature based on the Electromagnetic Articulograph. It could build 2D geometric models of lips and tongue for visualization. Feature extraction was divided in acoustic and articulatory features. In acoustic feature speech signal were dividing into acoustic frames and was extracted. Articulatory feature was done to calculate each EMA coil's displacement those was the Euclidean distance of the coil's position to its initial position. It could design 2D lips and tongue geometric models with B-spline curves and consist of front lip model and lateral model. After that apply GMM based method to the inversion mapping and done by acoustic-to-articulatory inversion mapping. The experimental result shows that the animations they synthesized were effective aids in helping people identifying vowels. They could control virtual articulatory models by multi speakers' data. In future work that expands system to syllable and continuous speech visualization for hearing aids.

The vision based technique and facial SEMG used for consonant and vowel identification respectively [1]. For silent speech detection SEMG used which sense visual and facial muscle activity. As consonant are easier to see and difficult to hear, therefore visual data is useful to classify consonant. In case of vowels the facial muscle activity useful, the reason is where the audio signal are weak or with noise visual information gives good results and vice-versa for vowels detection facial muscle movements are useful. Three steps- video recording for facial movement segmentation, visual feature extraction and classification are proposed. This work uses visemes to model visual speech, which implemented for facial animation applications by MPEG-4. The results show that the visual approach based on facial activity is suitable for consonant recognition. In this study the four facial muscles selected namely: Depressor anguli oris, Zygomaticus Major, Masseter and Mentalis. It also indicates that to identify the nine English consonants of the MPEG-4 the different patterns of facial movements can be used. In future the flexibility of regular conversation has to be designed.

As background study consist useful speech information, in addition study exhibit some desirable characteristics in comparison with acoustic signal for example resistant to acoustic noise. Therefore, if the speech information could be extracted with satisfactory accuracies, these signals can be used as secondary information sources for speech recognition in silent speech interface. Due to the limitations of data acquisition interface for the visual signal for example properly positioned camera or head tracking, consistent lighting conditions, only the facial SEMG signal were investigated in this dissertation. In the next chapter, the system development that used to extract the speech information for silent speech interface based on SEMG is discussed.

III. SYSTEM DEVELOPMENT

This section gives the goal to obtain accurate recognition of Labial Devanagari words from articulatory muscles movement, without using acoustic information. The Fig. 4 shows block diagram of system in below,

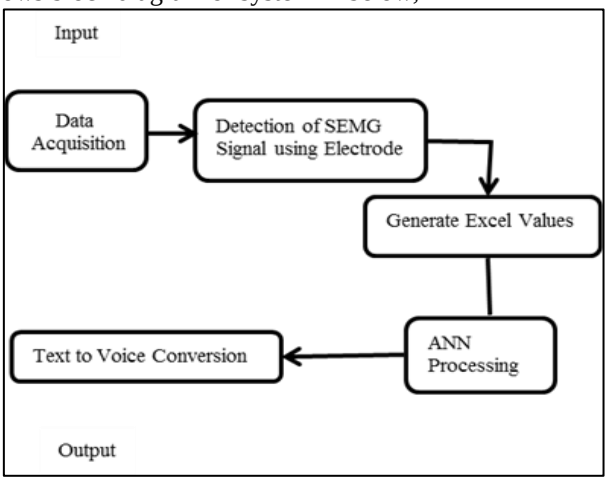

Fig. 4: Block Diagram of Proposed System

The above block diagram shows the experimental process of proposed system. The input signals of system are collected from articulatory muscles from SEMG using electrode. Collected input signal generates excel value which can be readable in MATLAB further the ANN is applied for processing. The processed signals are recognized and output is obtained in text to voice form.

A. Data Acquisition

One male speaker participated in the experiment. The EMG machine is used for SEMG recording which consist of 2 channel EMG configuration according to recommended recording guidelines. The electrodes were mounted on selected facial muscles. Before the recording commences following steps as, skin preparation, placement of electrode and removal of artifact are used.

B. Detection of SEMG signal

The detection of SEMG signal involves two hardware and software, since hardware collects signal from mouth muscles and software displays that signal. The detection of EMG signals from mouth muscle the device Quest 201 PolySomnoGraphy of Recorders & Medicare Systems Ltd. Company is used. PSG Machine includes both the Hardware and Software. PSG allows a wide variety of physical phenomena to be recorded simultaneously. The Software further helps to save multiple Records and provided with additional Filter Settings, Functions, Calculation Tools, Automatic analysis and Auto Report Generation.

As soon as the SEMG signal is acquired, the conditioning or processing of the signal is required to move further. In this processing of the signal refers to activity detection from the recorded SEMG signal. Activity detection is used to isolate the word said from the continuous SEMG stream.

C. Export Raw Data to Excel

Feature extraction is the process of reducing the size of the data to facilitate classification process. After the recoding process was completed, the raw EMG was transferd to MATLAB for further analysis in the form of excel values. To read SEMG signal into MATLAB, it is important to extract signal from PSG software. Therefore one application is provided into software, through which SEMG signal can able to convert into excel values. The exported Excel value for each sampled is applied to ANN processing in MATLAB.

D. ANN Processing

Recognition of EMG based speech features can be accomplished by supervised artificial neural network or statistical techniques. ANN learns to recognize the characteristic features of the data to classify the data efficiently and accurately. The model of neural network composed of many nerve cell is a classifier with self-learning ability and adaptively. The neural network includes ART network, Hopfield network, self-organizing network and BPNN. The input layer is the initial information imported in the network the hidden layer is determined by the input layer and its linked weights and the output layer is determined by the hidden layer and the linked weights of the output layer. The route of transmission is clearly visible is shown in Fig. 5.

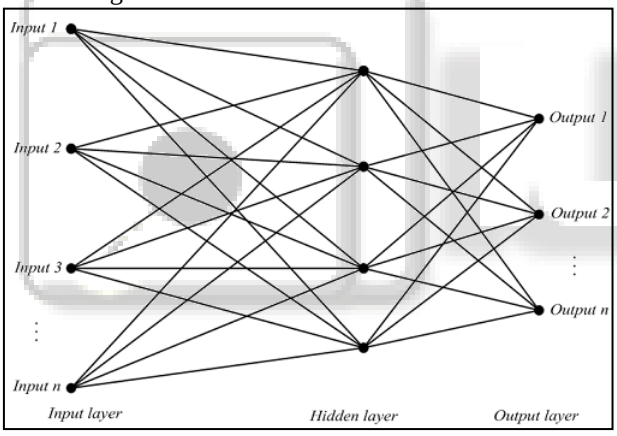

Fig. 5: The Structure of Neural Network

The transmission function of BP neuron, whose learning algorithm includes unsupervised learning and supervised learning, is nonlinearity. The supervised learning, the operation of network training of which is complex but valid, is shown in Fig. 6. In this the decomposing of SEMG is based on the neural network of the supervised learning.

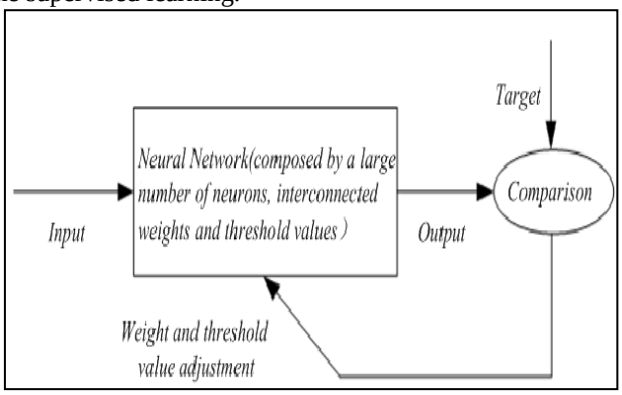

Fig. 6: Flow Chart for ANN

E. Text to Voice Conversion

The output of ANN is in the Text form, therefore text to voice converter is applied for getting final output.

IV. PERFORMANCE ANALYSIS

A single-speaker dataset consisting of four Devanagari Labial words was collected from healthy native Marathi speaker and used to evaluate the feasibility of proposed approach.

Table 2 lists the names and positions of the two articulators i.e sensors which are used for recognition.

Sr. No.	Articulators	Location
1	Orbicularis Oris Superior	Upper Lip
2	Orbicularis Oris Inferior	Lower Lip

Table 2: Names and Position of Articulators

The recognition of words is the focus of this result; it recognizes words from an unsegmented sequence of articulatory motion by analyzing the probability graph returned from the trained model. The correct word should have higher matching probability than any other word at the time it occurs. However the length of each word is unknown, therefore it needs to determine the location for each word.

The performance algorithm was measured in terms of its recognition accuracy. In each execution, a prediction was deemed correct only when the predicted word was correct. Recognition accuracy and processing time were used to evaluate the performance of the word recognition algorithm. A word prediction is correct if the expected word is identified within half a second of its actual occurrence of time. That is, both missing values and wrongly predicted occurrence time are considered as errors.

The subject obtained word recognition high accuracy. The high accuracy and relatively short delay demonstrated the feasibility of SSIs based on Electromyography. Although the experiment was conducted offline i.e. data were collected before the analysis, the normalization approach can be easily integrated into online silent speech interface. The approach is also advantageous for other online applications because it does not require prerecorded data.

Furthermore, in this proof of concept design, the vocabulary was limited to a small set of words. The additional work needed to test the feasibility of open vocabulary recognition, which will be more usable for people after laryngectomy.

V. CONCLUSION

SSI will enable to communicate without an audible acoustic signal. The acoustic speech recognition is substituted by silent speech recognition, in which EMG signals are detects from the surface. The word recognition system uses EMG technology to sense the labial speech signal, the signals converted to speech again using prerecorded commands and classification of words from neural network. Therefore EMG based speech yields good improvements in SSI.

The experimental results shows the potential of word recognition algorithm for building an articulation based silent speech interface, which can be used in command and control system using silent speech and may even enable voiceless patient to produce synthetic speech

using articulatory movement. The result is revealing high word recognition accuracy and short playback. Although the current results are encourages for future work as larger vocabulary including phonemes from a larger number of subjects with different genders and dialects are necessary to explore the limits of the current approach and built a real time speech recognition system for SSI.

REFERENCES

[1] Wai Chee Yau, Sridhar Poosapadi Arjunan and Dinesh Kant Kumar "Classification of Voiceless Speech Using Facial Muscle Activity and Vision based Techniques" Australia.
[2] Eduardo Lopez-Larraz, Oscar M. Mozos, Javier M. Antelis, Javier Minguez, "Syllable-Based Speech Recognition Using EMG" 32nd Annual International Conference of the IEEE EMBS, September 2010 IEEE, PP. 4699-4702.
[3] Peter Konrad, "The ABC of EMG", A Practical Introduction to Kinesiological Electromyography, Version 1.0 April 2005.
[4] Uwe Windhorst, "Modern Techniques in Neuroscience Research", Springer, 1999.
[5] Sanjay Kumar, Dinesh Kant Kumar, Melaku Alemu, and Mark Burry, "EMG Based Voice Recognition".ISSNIP 2004 IEEE, PP. 593-598.
[6] Chan, A., Englehart, K., Hudgins, B., Lovely, D., 2001.Myoelectric Signals to Augment Speech Recognition. Medical and Biological Engineering and Computing 39, PP. 500 – 506.
[7] Jun Wang, Arvind Balasubramanian, Luis Mojica de la Vega, Jordan R. Green Ashok Sama, Balakrishnan Prabhakaran, "Word Recognition from Continuous Articulatory Movement Time-Series Data using Symbolic Representations" SLPAT 2013, pages 119–127, Grenoble, France, 21–22 August, 2013.
[8] Michael Wand, Christopher Schulte, Matthias Janke, Tanja Schultz, "Array-based Electromyographic Silent Speech Interface" Cognitive Systems Lab, Karlsruhe Institute of Technology, Karlsruhe, Germany.
[9] Szu-Chen Stan Jou, Tanja Schultz, and Alex Waibel, "Continuous Electromyographic Speech Recognition with A Multi-Stream Decoding Architecture" ICASSP-88,1988 10.1109/ICASSP.2007.
[10] Szu-Chen Jou, Tanja Schultz, Matthias Walliczek, Florian Kraft, and Alex Waibel, "Towards Continuous Speech Recognition Using Surface Electromyography" INTERSPEECH 2006- ICSLP.
[11] Matthias Janke, Michael Wand, and Tanja Schultz, "A Spectral Mapping Method for EMG-based Recognition of Silent Speech" Cognitive Systems Lab, Karlsruhe Institute of Technology, Karlsruhe, Germany.
[12] Jun Wang, Ashok Samal, Jordan R. Green, "Across-speaker Articulatory Normalization for Speaker-independent Silent Speech Recognition", INTERSPEECH 201414-18 September 2014, Singapore.
[13] Lena Maier-Hein, Florian Metze, Tanja Schultz, and Alex Waibel, "Session Independent Non-Audible Speech Recognition Using Surface Electromyography".0-7803-9479-8/05, IEEE ASRU 2005.
[14] Hao Li, Minghao Yang, Jianhua Tao, "Speaker-Independent Lips and Tongue Visualization of Vowels", 978-1-4799-0356-6/13 IEEE, ICASSP 2013.

