Different Speech Recognition EMG based Approaches for SSI

Geeta N. Sonawane[1] Mr. C. S. Patil[2] Mrs. A.N.Shewale[3] Mr. R. R. Karhe[4] Gunjan G. Gujarathi[5]
[1]Research Student [2]Head [3,4]Associate Professor [5]Assistant Professor
[1,2,3,4,5]Department of Electronics & Telecommunication & Engineering
[1,2,3,4,5]SGDCOE, NMU, Maharashtra, India

Abstract— This paper reflects the different sampled data used in speech recognition for SSI. The EMG is an electrical potential generated by movements associated with speech sound production in glottal region, jaw, tongue, soft palate, lips and other areas. This paper provides up-to date review of current approaches based on speech recognition EMG based speech recognition. The techniques are helpful to enable SSI, when EMG signals are recorded for those people who only articulate speech without producing any sound. The SSI have mostly depends on sampled data for facial and visual modalities to reduce the negative effect in speaker variation and recognition accuracy. The outline of discussion that it will need to improve in large training dataset, used of different languages based model, removal of noise artifacts.

Key words: Articulatory Muscles, Electromyography-Based Speech Recognition, Modalities, SEMG, Silent Speech Interface

I. INTRODUCTION

One of the electronic systems that enable to communicate by speech without an audible acoustic signal is Silent Speech Interface [1]. In contravention of their 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. In voice communication speech needs to be clearly audible without masked and includes lack of robustness in noisy environments, privacy issues, disturbance for bystanders, and exclusion of speech disabled people. In perverse environmental conditions as, in restaurants, cars, or trains the speech recognition performance reduce considerable.

Using audible speech signal a confidential conversation with or through a device becomes impossible [2]. 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, for example 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.

Alternatives for Communication that are both private and not dependent on production of audible signals are valuable [3]. Presently, there are limited treatment options for: esophageal speech which is difficult to learn, electro larynxa mechanical device resulting in a robotic-like voice, and augmented and alternative communication (AAC) devices for example, text-to-speech synthesizers operated with keyboards.

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 [4]. 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. Furthermore, two techniques for data collection of EMG signal: surface method and intramuscular method [5]. In this paper focused on surface method for data collection, since compared with intramuscular EMG signal, surface EMG signal has non-invasive, fast, co-ordinated and other characteristics, becoming the mainstream in many areas in recent years. Common EMG signals strength can range from uV to few mV. These surface EMG signals normally are of low amplitude. The generalized block diagram to process with SEMG signals is as given Fig. 1;

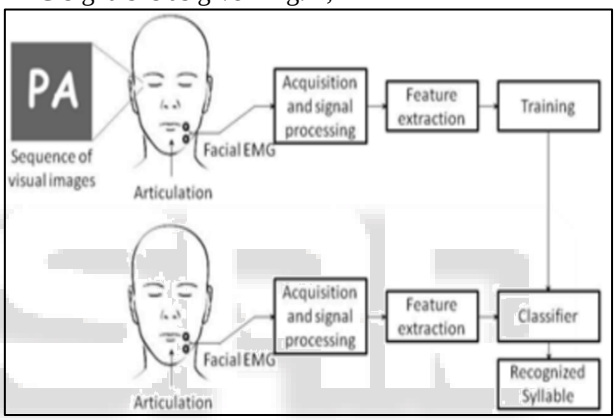

Fig. 1: The generalized block diagram to process with SEMG signals [6]

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 [7]. This paper illustrates the speech recognition techniques. Paper presents feature extraction methods in section II and section III concludes the paper.

II. RELATED WORK

Generally for non-acoustic speech recognition two modalities are used namely visual and facial muscles activity based on SEMG for identification of silent speech. This review describes the facial and visual approaches. 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 Modalities

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 [8]. 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 workstation 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 [6], 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 [6]

Author focused on syllable of Spanish language based on EMG signal recorded in facial muscles [6]. 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.

The prime requirement of word recognition for SSI discussed here in [9]. 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 SAX 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.

To recognize the word, one ore 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. Author proposed the first application of the new array technology [11]. They show that the recognition result improved by Independent Component Analysis (ICA). 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.

EMG continuous speech recognition is a system which uses the information from EMG articulatory feature suggested in [12]. 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. To extract the signals from articulatory muscles the six electrode pairs are positioned as shown in Fig. 2;

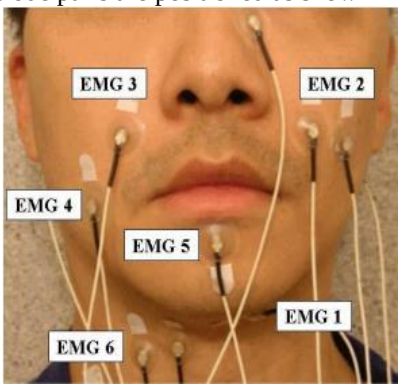

Fig. 2: EMG Electrode Positioning [12]

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 Jouhad taken research for continuous EMG speech recognition system on normal audible speech [13]. 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 [14]. 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 levatoranguli oris, the zygomaticus major, the platysma, the anterior belly of the digastrics and the tongue. The feature extractions were based on time-domain features and normalize the frame. They use cross model inialization 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 they introduced Speaker independent speech recognition method [15]. 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;

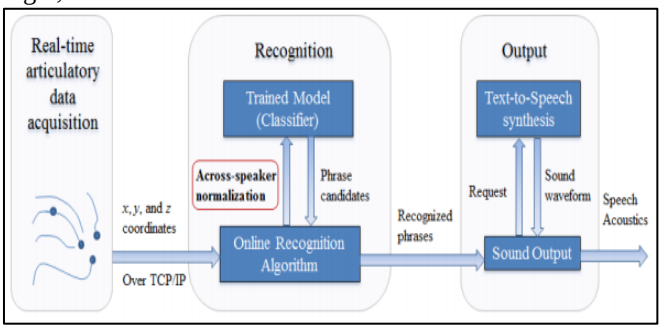

Fig. 3: Conceptual design of the speaker-independent silent speech interface [15].

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 in [16]. 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 paly vital role in human interaction, therefor visual and SEMG signals are selected for HCI applications. The author had focused on the visualization of articulators from acoustic signal frame by frame in [17]. 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 by SPTK. Articulatory feature was done by DRD 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, therefor 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 are 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. The muscles location for electrodes placement are shown in Fig. 4.

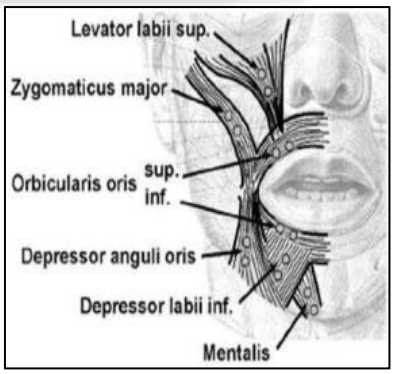

Fig. 4: Topographical location of facial muscles [1]

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.

III. Future Challenges

In this paper the brief literature review of feature extraction techniques are discussed. Many researches had done valuable work to enhance SSI applications. EMG based SSI has been active research topic since last two years, but comparisons of different technique for speech processing still difficult. The performance of speech recognition also plays vital role in statistical language model. Perhaps, some discrepancies remain to focus,

- Improving speech based multistream pronunciation model.
- Implementing different language based model.
- Improving accuracy, repeatability and reproducibility of experimental data set.
- Removal of artifact during data acquisition.
- Extension of dataset to a larger set of vowels, words, sentence and digits.
- Built a real time speech recognition system for SSI.

IV. 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. In this article, the facial and visual modalities described where the results are satisfactory on their independent level. With this approach, EMG based SSI effectively works for vowels, words, sentence, digits in facial and visual modalities too. Therefore, EMG based speech yields good improvements in SSI.

Acknowledgement

Authors would like to express gratitude to Prof. C.S.Patil, Head of the Electronics and Telecommunication Department for his Guidance and support in this work. The authors also like to thankful to the Principal Dr. A.J. Patil of SGDCOE, Jalgaon, Mrs. A.N.Shewale and Mr. R.R.Karhe for being a constant source of inspiration.

References

[1] WaiChee Yau, Sridhar Poosapadi Arjunan and Dinesh Kant Kumar "Classification of Voiceless Speech Using Facial Muscle Activity and Vision based Techniques" Australia

[2] Matthias Janke, Michael Wand, Keigo Nakamura, Tanja Schultz, "Further Investigations on EMG-To-Speech Conversion", 978-1-4673-0046-9/12/IEEE ICASSP 2012.

[3] Jun Wang, Jordan R. Green, Ashok Samal "Vowel Recognition from Continuous Articulatory Movements for Speaker Dependent Applications", 10.1109/ICSPCS.2010.5709716,ICSPCS.

[4] Ms. Rutuja U. Bachche, Prof. R.T.Patil, "EMG Signal Feature Extraction for Designing the Calf Stimulator", IJARCETVolume 4 Issue 6, June Volume 4 Issue 6, June 2015 ISSN: 2278 – 1323

[5] Bo You, Shoutong Tao, Yi Liu and Hanqing Zhao, "The Verification of Physiological Model of SEMG Based on Wavelet Decomposition", IJSIPVol.8, No.5 (2015), PP. 341-352.

[6] 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(978-1-4244-4124-2/10/$25.00 ©2010 IEEE) 4699-4702

[7] Michael Wand and Tanja Schultz, "Analysis of Phone Confusion in EMG-based Speech Recognition", 978-1-4577-0539-7/11 IEEE ICASSP 2011.

[8] Sanjay Kumar, Dinesh Kant Kumar, Melaku Alemu, and Mark Burry, "EMG Based Voice Recognition". ISSNIP 2004 IEEE, pp. 593-598.
[9] 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.
[10] Chuck Jorgensen, Diana D. Lee, and Shane Agabon, "Sub Auditory Speech Recognition Based on EMG/EPG Signals"
[11] Michael Wand, Christopher Schulte, Matthias Janke, Tanja Schultz, "Array-based Electromyographic Silent Speech Interface "Cognitive Systems Lab, Karlsruhe Institute of Technology, Karlsruhe, Germany.
[12] Szu-Chen Stan Jou, Tanja Schultz, and Alex Waibel, "Continuous Electromyographic Speech Recognition with A Multi-Stream Decoding Architecture"ICASSP-88.,
[13] Szu-Chen Jou, Tanja Schultz, Matthias Walliczek, Florian Kraft, and Alex Waibel, "Towards Continuous Speech Recognition Using Surface Electromyography" INTERSPEECH 2006- ICSLP.
[14] 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.
[15] Jun Wang, Ashok Samal, Jordan R. Green, "Across-speaker Articulatory Normalization for Speaker-independent Silent Speech Recognition", Interspeech 201414-18 September 2014, Singapore.
[16] 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.
[17] Hao Li, Minghao Yang, Jianhua Tao, "Speaker-Independent Lips and Tongue Visualization of Vowels", 978-1-4799-0356-6/13 IEEE ICASSP 2013.

