Subvocal Speech Recognition System based on EMG Signals

Yukti Bandi

Assistant professor
DJ Sanghvi College of Engineering, Mumbai

RiddhiSangani , Aayush Shah,

AmitPandey, ArunVaria

UG Students,
DJ Sanghvi College of Engineering , Mumbai

ABSTRACT

This paper presents results of electromyography (EMG) speech recognition which captures the electric potentials that are generated by the human articulatory muscles. EMG speech recognition holds promise for mitigating the effects of high acoustic noise on speech intelligibility in communication systems. Few words have been collected from EMG from a male subject, speaking normally and sub vocally. The collected signals are then required to be filtered and transformed into features using Wavelet Packet and statistical windowing techniques. Finally, the concept of neural network with back propagation method has been used for classification of data.Using windowed signals and the trained neural network an arduino operated bot was controlled as an application to demonstrate the future scope of the paper. The success rate was 73%.

Keywords

EMG, sub vocal speech, neural network, electromyography

1. INTRODUCTION

Human speech communication typically takes place in complex acoustic backgrounds with environmental sound sources, competing voices, and ambient noise. The presence of such noise makes it rather difficult for the human speech to remain robust. In many applications minimum noise with optimum output is essential. For the purpose of better communication over a range, transmitters and receivers with high noise filtering capabilities are developed [1][4]. Researchers are working extensively on improving the bioelectric signalling. The goal being to completely eliminate the noise generated during transmission and reception of signals to further the advancement of human-computer interaction.

Sub vocal speech is silent, or sub-auditory, speech, such as when a person silently reads or talks to himself [13]. The electromyograms of each speaker are different from each other. So the signals produced vary for every individual. The accuracy of the signal can be improved by enhancing pattern recognition characteristics. Sub vocal signals are gathered non-invasively by attaching a pair of electrodes on the throat and, without opening the mouthor uttering a word. words are recognized by a computer. The signals obtained froelectromyography' are used as bioelectric signal. The EMG signals evaluate and record the electrical activity producedelectrically or neurologically activated. The signals can be analysed to detect medical abnormalities, activation level, or recruitment order or to analyze the biomechanics of human movement [13].Surface EMG (sEMG) assesses muscle function by recording muscle activity from the surface above the muscle on the skin. For that purpose,more than one electrode is needed because EMG recordings display the potential difference (voltage difference) between two separate electrodes. Once the signal is acquired from the sEMG, feature extraction will be carried out using time-frequency representations using Discrete Wavelet Transform.

This paper illustrates the techniques and steps employed in the acquisition, analysis, and how the signal will be processed and classified of sub-vocal speech.

2. METHODOLOGY

2.1 Data Acquisition

The EMG signals of sub-vocal speech were acquired using surface electrodes of Silver Chloride (AgCl) in bipolar configuration [3] .These electrodes were kept on the upper part on the right side and lower part on the left side under the throat (see Fig 1). Using isopropyl alcohol at 70 % reduction in impedance of the skin was achieved. (to compensate for the low amplitude of sub-vocal signals), the EMG signals were amplified by a factor of 1000.

2.1.1 Noise Reduction

In the acquisition system to reduce noise and to attenuate frequencies which are not part of the EMG signal, a band-pass chebyshev filter of order 4 is used, formed by a high-pass filter of 25 Hz and a low pass filter of 450 Hz. A preamplifier of gain 100 has been implemented and a post amplifier accompanied along the filter to obtain gain of about 1000.

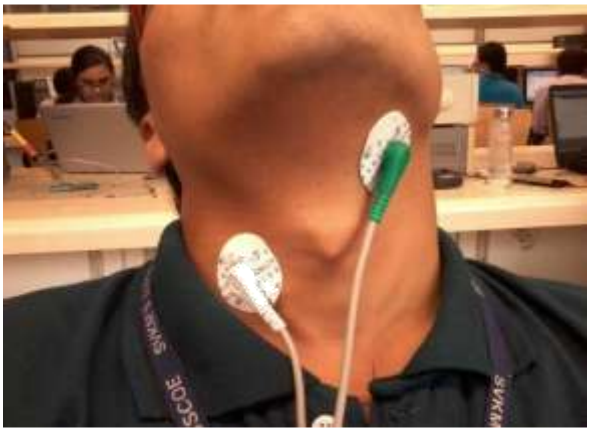

Fig 1: Locations of electrodes on the subject

2.1.2 Transfer of Data to PC

After the filtering process is completed, the filtered signal is then required to be transferred to the computer for further analysis. Various methods have been followed to transfer the data to pc, they are as follows:

- Using A/D convertor of PSOC4, using it at 8 bit resolution and in differential mode with sampling rate of 166666sps. But when tested it with the filtered emg signal, it didn't give satisfactory results.
- Another method proposed was to serially get the signal through the dso in MATLAB 7.11, but when analyzed, it also didn't give satisfactory results.
- But finally the method which gave satisfactory results was using a mono jack (one side male and other side striped to give the signal). This mono jack was connected to the microphone jack in pc. Sound recorder of the pc was used to record sessions of few seconds. The recorded signal was in .wma format which was converted to .wav file for using it with MATLAB 7.11
- A timer was set for 10, 15, 20 seconds for different sessions. Following figures shows the recorded signal read and plotted in MATLAB 7.11.See figures 2, 3, 4. In these figures X axis represents no. of samples and Y axis represents amplitude.

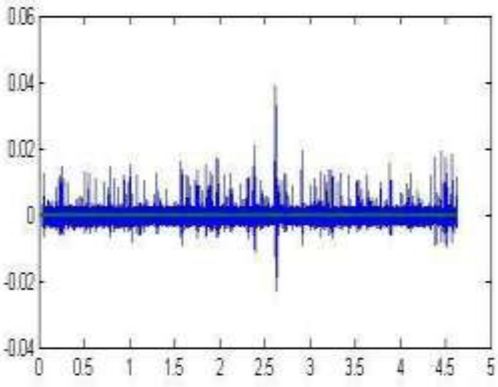

Fig 2 : Recorded signal for 10 seconds when subject was completely silent

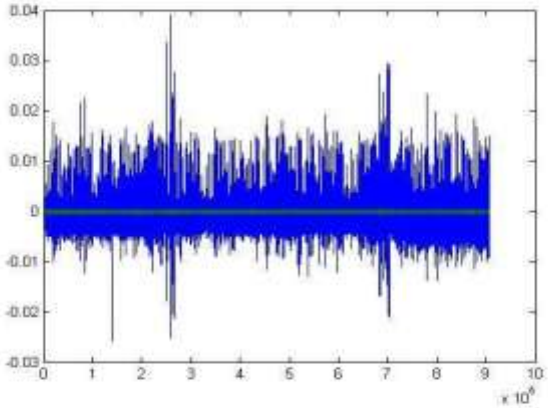

Fig 3 : .Recorded signal for 20 seconds when subject said forward at 5th second and 15th second

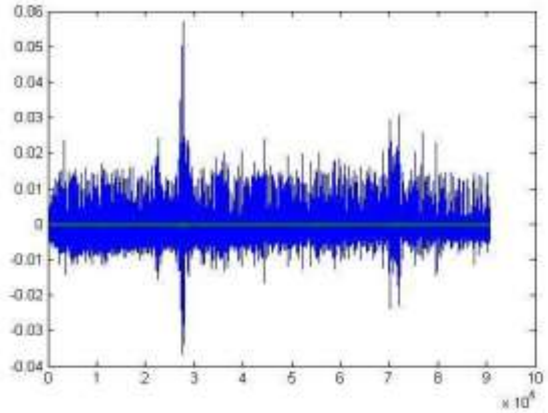

Fig 4 :Recorded signal for 20 seconds when subject said forward at 5th second and reverse at 15th second

2.2 Signal Conditioning

As soon as the sEMG signal is acquired, the conditioning or processing of the signal is required to move further. Here processing of the signal refers to activity detection from the recorded emg signal. Activity detection is used to isolate the word said from the continuous emg stream [7]. There are various ways for doing that such as:

- Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold [10]
- Voice activity detection using higher order statistics [11].
- Statistical voice activity detection using a multiple observation likelihood ratio test [12].
- Sudden change in energy detection using energy sensitive windows.

The research carried out by NASA reveals that they have used last method mentioned above in their basic model and rest methods were mentioned as highly sophisticated methods which were reserved for future work.

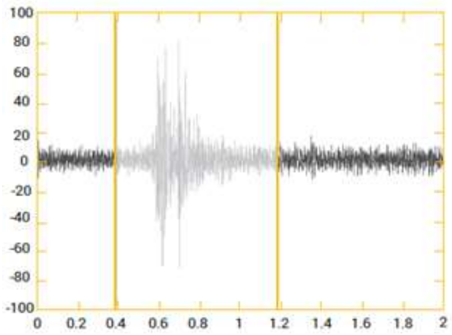

Fig 5 : Active zone of the sub-vocal signal for the word "Forward"

The basic idea behind defining the window is that it can be defined in such a way that they can detect the sudden change in energy within their domain. Sudden change in energy refers to the presence of a word, as noises are of lower amplitude and also almost of same amplitude among them. As soon as the word is uttered the energy for that duration rises as shown in Fig 4. The work of the window is to detect such energy changes, so that the noises can be eliminated and only the word can be extracted (see fig 5).

A small window size was selected using energy extraction and then locating maximum energy to locate the active region ,the size of window should be small but not that smaller that the active region gets divided in two halves. The signal energy for windowing is defined as [6]:

$$E_n = \sum_{i=1}^{W}(X_{(n-1)}W+i) \qquad (1)$$

Energy extraction was done by root sum of squares of samples in segments equivalent to 100ms which were then averaged using sliding window and consider 10 energy samples at a time. This gave an active window of one second which was detected by taking maximum value of averaged signal and then spreading window in forward and backward direction. Energy of signal is given by [6] :

$$\upsilon[n] = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1} x_{n^2}\,(i)} \qquad (2)$$

After locating the active region hard thresholding was used to remove lower energy signal hence, reduce effect of noise. Threshold was found using [6]:

$$U = (0.15)\ E_{max} \qquad (3)$$

Where U is the threshold value and E_{max} is the peak energy. After locating the active area the windowed signal is then filtered using discrete wavelet transform (DWT) which is expressed by equation no. 4[6]:

$$\Psi_{jk}\,[n] = 2^{\frac{j}{2}}\Psi\left(2^{j}\,\mathrm{n}-\mathrm{k}\right)\mathrm{j},\mathrm{k}\epsilon\mathrm{Z} \qquad (4)$$

Where j is the scaling factor, k is the translational parameter and Ψ is the wavelet function. In this process the signal is divided into four parts using mother wavelet daubechies [3]. The threshold for the filter can be defined as [6]:

$$\mathrm{U}f = \sqrt{2 * \log n} \qquad (5)$$

Where n represents the no. of samples Uf represents the threshold value. As the threshold is calculated, hard thresholding to windowed signal was done. The hard thresholding method is defined as[6]

$$f(x) = \begin{cases} x\,, & |x| > U_f \\ 0\,, & |x| < U_f \end{cases} (6)$$

Thus, after filtering the signal is reconstructed using inverse discrete wavelet transform given by [6]:

$$x[n] = \sum_{j\in Z}\sum_{k\in Z} C[j,k].\psi_{j,k}[n] \qquad (7)$$

Here x[n] is the reconstructed filtered signal, c represents the thresholding coefficients and Ψ is wavelet bias.

2.3 Feature Extraction

Feature extraction is the process of reducing the size of the data to facilitate classification process [7]. This process can be carried out in two ways.

In first method feature extraction is carried out by using discrete wavelet packet transform (DWPT) [5][6].

DWPT decomposes the averaging coefficients and the detail coefficients forming a tree structure [1].

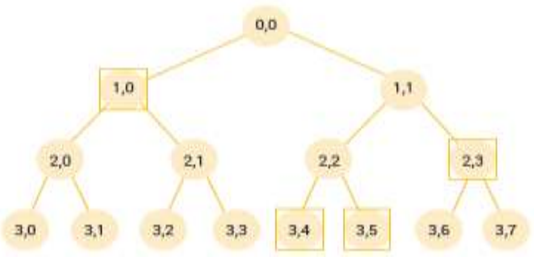

Fig 6 : Selecting the Best DWPT Basis [6]

To select the optimum basis function or coefficients that best represent the sub-vocal signal, a cost function that measures the level of information is used. In this case, Shannon's entropy was used as the cost function, namely [3][6]:

$$H(q) = \sum_{k}|q(k)|\log\frac{1}{q(k)} \qquad (8)$$

Where H represents the Shannon entropy, q(k) is normalized energy of the wavelet coefficients. Once the cost function is found next step is to choose optimum bias by means of parameters suggested in[6].

Using these results of DWPT, the patterns were extracted by means of statistical methods such as root mean square (RMS) etc. Second method uses the discrete wavelet transform which approximates the values of the signal and reduces the size of the signal. On application of DWT the size of the signal reduces to half. Using repeated application of DWT the size can be reduced to adequate level. Using statistical methods and DWT process of principal component analysis (PCA) was used to reduce the size of data and it is found that the most of the information of the data is contained in first few elements of the coefficients hence only these coefficients are used for classification [3].

2.4 Classification

Classification of the data is done by using neural network, a multilayer perceptron neural network with supervised back propagation. The network was first trained using test data with no. of nodes equal to selected coefficients as input nodes, a layer of hidden layer and one output node. Using this trained network the signals were tested and efficiency of each word was calculated.

3. EXPERIMENTAL RESULTS

In this section, results of the recognition system and the acquired digitized outputs and feature extracted signals have been shown.

3.1 Classification Effectiveness

Classification of the signals was done by Feed Forward Back Propagation having 4 hidden layers and 361 input neurons with 1 output neuron. The training set consisted of 100 samples with 50 samples of each word. The database was constructed by capturing signals from a 22-year-old male person in various recording sessions, with conventional noise conditions.

Table 1 shows the results of the classification phase, it can be seen that effectiveness of the algorithm varies depending on the sub-vocal signal being classified, with an average accuracy of 75%.

Table 1: Effectiveness of the classification process

WORD	EFFICIENCY (%)
Forward	74.4
Reverse	72.5

3.2 Data Acquisition

The signal was acquired using the designed circuit and then digitized using the sound card of computer. The signals were recorded for span of 10 seconds and then observed on Matlab software. The samples were taken for both vocal speech and silent lipsing (see figure 7,8).

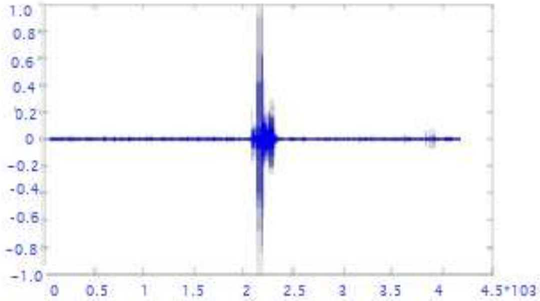

Fig 7: Sampled vocal forward signal

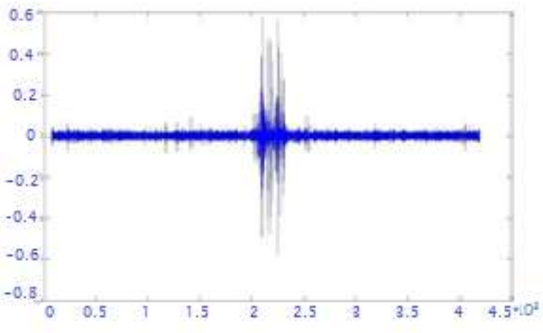

Fig 8 : Sampled lipsing reverse signal

3.3 Signal Conditioning

The recorded signal was then windowed to find the active region. The process of windowing was done using energy extraction and it was observed that the signal amplitude above 0.6 was detected in the active region. The window of 1 second was selected as the active region.

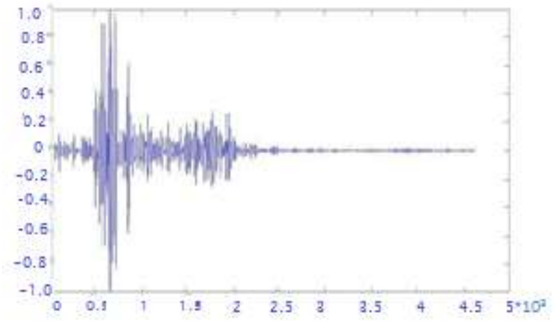

Fig 9 : Windowed signal for forward signal

3.4 Feature Extraction

The windowed signal was passed through Discrete Wavelet Transform using 'db1' harr transform to remove excess noise and reduce no. of samples. Fig 10 shows the result after 7th transform. The no. of reduced samples were 361 samples per signal.

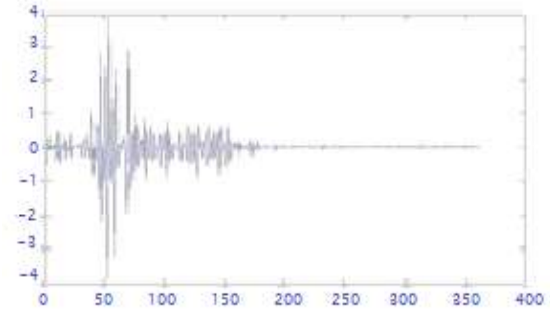

Fig 10 : Signal after 7th transform

3.5 Robot Control

Using these signals and the trained neural network an arduino operated bot was controlled as an application to demonstrate the future scope of the topic. As well the signals were converted to speech again using prerecorded commands and classification of word from neural network. Hence shows future scope of the topic in bio medical field.

4. CONCLUSION

The subvocal speech recognition system uses EMG technology to sense the vocal speech signal and controls the device in real time. The precise control & accuracy can be achieved by accurate design of pre-amplifier, post amplifier, band pass filter and proper training of neural network.

5. FUTURE SCOPE

The sub vocal speech recognition system is still under research by various research institutes. It is developed for the small words (or set of words), not for the continuous communication. Thus it can be implemented for continuous communication by training neural network precisely. The sub-vocal signals will be transmitted wirelessly to the processing real-time-recognition system for confidential military communication in noisy acoustic environment. The feature-extraction (and classification) algorithm will be improved to obtain a higher accuracy and precise control.

6. REFERENCES

[1] Chuck Jorgensen, Diana D. Lee and Shane Agabon, "Sub Auditory Speech Recognition Based on EMG Signals,"Proceedingsof the International Joint Conference on Neural Networks (IJCNN), IEEE, vol. 4, 2003, pp. 3128–3133.

[2] Chuck Jorgensen and Kim Binsted, "Web Browser

[3] Control Using EMG Based Sub vocal Speech

[4] Recognition, "Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS), IEEE, 2005, pp. 294c.1–294c.8.

[5] Luis Enrique Mendoza, Jesús Peña Rodríguez, Jairo Lenin Ramón Valencia. "Electro-myographic patterns of sub-vocal Speech: Records and classification" Research Group of GIBUP University The Pamplona, *Colombia.November 29, 2013.*

[6] RatnakarMadan, Prof. Sunil Kr. Singh, and NitishaJain," Signal Filtering Using Discrete Wavelet Transform", published in International Journal of Recent Trends in Engineering, Vol 2, No. 3, November 2009.

[7] Mark C. Goñi and Alexander P. de la Hoz, "Analysis of Biomedical Signals Using Wavelet Transform " ContestStudent Jobs EST , National University of San Martin,Argentina , 2005.

[8] Dora María Ballesteros Larrotta, "Application of Discrete Wavelet Transform Filtering bioelectric signals," Threshold Scientific, Manuela Beltran University Foundation, Bogotá, Colombia, pp. 92-98 ,Dic. 2004.

[9] Bradley J. Betts, Charles Jorgensen, "Small Vocabulary Recognition Using Surface Electromyography in an Acoustically Harsh Environment", National Aeronautics and Space Administration (NASA), Ames Research Center Moffett Field, California, 94035-1000, November 2005.

[10] FA Sepulveda, "Extraction of Speech Signals Parameters techniques using Time-Frequency Analysis , " National University of Colombia , Manizales, Colombia , 2004.

[11] Muhammad Zahak Jamal "Signal Acquisition Using Surface EMG and Circuit Design Considerations for Robotic Prosthesis".Intech 2012.

[12] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Transactionson Speech and Audio Processing*, to appear, pp. 1–13.

[13] K. Li, M.N.S. Swamy, and M.O. Ahmad, "An improved voice activity detection using higher order statistics," *IEEE Transactions on Speech and Audio Processing,*vol.13, no. 5, 2005, pp. 965–974.

[14] .J. Ramírez et al., "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEESignal Processing Letters*, vol. 12, no. 10, 2005, pp.689–692.