Proceedings of the 8th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'22)
Prague, Czech Republic - July 28- 30, 2022
Paper No. ICBES 141
DOI: 10.11159/icbes22.141

Myo-Speech: A System for Recognizing Word Utterances of the Speech Impaired

Aya S. Al-Mowafy, Mona M. Abd El-Aty, Ahmed A. Morsy
Biomedical Engineering and Systems Department, Cairo University
Giza ,12613, Egypt
aya.saeedmwafy@gmail.com; mona.mohamed.aty@eng1.cu.edu.eg; amorsy@eng1.cu.edu.eg

Abstract **-** We report a new method for identification of unintelligible vocal sounds obtained from two speech impaired volunteers. The ultimate goal of this research is to generate some speech production capabilities for persons with speech impairment by reissuing intelligible versions of words corresponding to the vocal sounds they generate in their typically unsuccessful efforts to produce spoken words. The ability to generate understandable spoken words can have a major positive impact on the quality of life of speech impairment communities around the world. To acquire classifiable signals corresponding to the produced vocal sounds we did not use a conventional microphone as it is not immune to background noise. Rather, we used surface electromyography (sEMG) signals obtained from the sternocleidomastoid muscle in proximity of the vocal cords, with the hypothesis that there should be consistent correlation between the intended words and their myo signals. A preliminary dataset consisting of three Arabic words was acquired from two subjects (one female and one male) in a lab-controlled environment. Average classification accuracy was about 82% using standard machine learning techniques (KNN, NN, and SVM). Preliminary results indicate that classification of this limited vocabulary dataset is feasible with reasonable accuracy to motivate future work involving more subjects and larger datasets.

Keywords**:** Speech impairments, sEMG, Machine learning, Vocal sounds.

1. Introduction

Communication is one of the most important skills for human beings. It is the way through which they send and receive information. Typically caused by at-birth or early childhood hearing loss, speech impairment is one major issue that hinders communication. According to the WHO, more than 5% of the world population suffer from a certain degree of hearing loss [1]. The WHO estimates that around 700 million people will be having disabling hearing loss by 2050 [2]. Around 80% of people with hearing loss live in developing countries. Disabling hearing loss is usually associated with speech impairment, especially in developing countries [2]. Disabling hearing loss is usually associated with speech impairment, especially in developing countries, as children are often not schooled [3]. Humans with disabling hearing loss have limited communication abilities, which compromises not just their quality of life but also the value that they can add to their societies. The use of sign language may date back to the fourteenth century [4]. Although thousands of hearing and speech impaired persons rely on sign language around the world, its use still does not form a universal solution for efficient societal integration of the speech and hearing impaired. Recent research used artificial neural networks to decode sign language into some form that can be more widely understood [5].

Some recent research also focused on decoding facial and neck surface Electromyography (s-EMG) into the intended speech for non-speech impaired people using machine learning and deep learning [6,7,8,9]. The goal of this paper is to use data from speech-impaired subjects representing word utterances typically not understood because they are unintelligible. We acquired surface Electromyography (sEMG) signals corresponding to these word utterances to avoid challenges posed by the background noise if a microphone was used. The goal of the research is to use standard machine learning techniques to recognize these word utterances. By reproducing intelligible versions of these utterances, it would be possible in the future to give some speech production capabilities to the speech impaired persons

2. Material and Methods

2.1. Corpus Design and Participants

This preliminary study was carried out using the three Arabic words whose English meanings are Book, Door, and Metro for the first subject. The second word was replaced with a person's name (WASEEM) for the second subject. These words were selected based on the following criteria as applied to the Arabic Language: (1) different length, (2) different phonemes (i.e., different places of articulation; alveolar, plate, lips, teeth, etc.); and (3) inclusion of proper names.

The data was collected from two speech impaired subjects with severe hearing loss (1 female, age 24; and 1 male, age 25). Communication with the two subjects during the experiment was facilitated by a professional sign language translator. The translator explained the nature of the research to the subjects who sign an informed consent to participate in the experiments.

2.2. Experiment Protocol and Data Acquisition

Each subject was asked to relax for 5 seconds then pronounce a single word then relax for 5 seconds. This sequence was repeated 50 times for each of the three words. In total, each subject gave 150 records, out of which 120 records were used for training and 30 for testing.

The data was captured using sEMG kit (PLUX – Wireless Biosignals, Lisbon, Portugal) with OpenSignals (r)evolution software. The kit has a sampling rate of 1 KHz and 10-bit resolution ADC. The electrode was placed on the neck area on the sternocleidomastoid muscle as shown in Figure 1.

Figure 1. sEMG electrode placement.

2.3. Data analysis

2.3.1 Denoising

The input signals were pre-processed using a 2nd order bandpass filter [60-180 Hz] to remove the DC component and to decrease the noise [10,11] as shown in Figure 2.

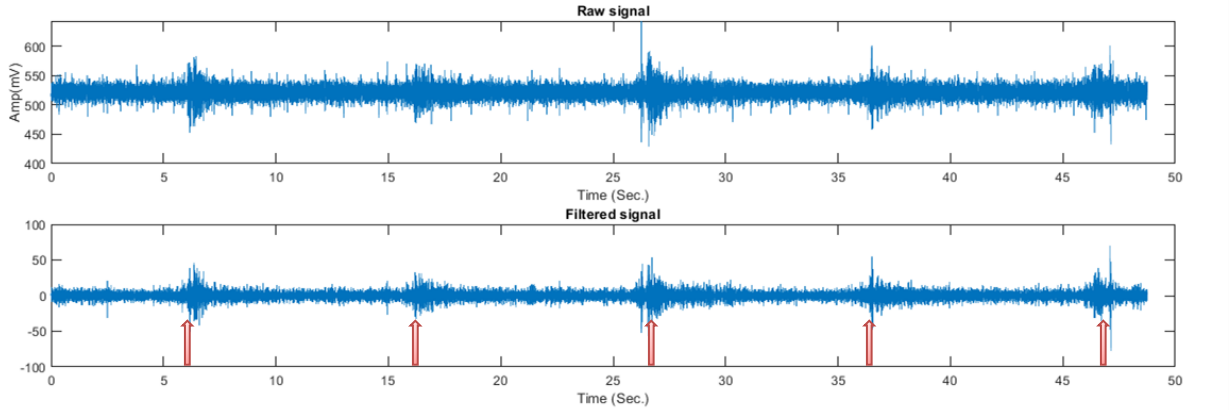

Figure 2. A sample of the acquired sEMG signals before and after filtration. Arrows indicate word utterances.

2.3.2 Feature Extraction

Ten time-domain features were extracted from the filtered signal as per [12]. These features are:

$$MAV = \frac{1}{N}\sum_{n=1}^{N} |x_n| \tag{1}$$

Where N denotes the length of the signal and x_n represents the sEMG signal in a segment.

$$SSI = \sum_{n=1}^{N} |x_n|^2 \tag{2}$$

$$VAR = \frac{1}{N-1}\sum_{n=1}^{N} x_n^2 \tag{3}$$

$$RMS = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2} \tag{4}$$

$$WL = \sum_{n=1}^{N-1} |x_{n+1} - xn| \tag{5}$$

$$Zc = \sum_{n=1}^{N-1} (x_{n^*}x_{n+1}) \cap |x_n - x_{n+1}| \geq threshold] \tag{6}$$

Where $sgn(x) = \{1 \quad , \quad if\ x \geq threshold\ 0 \quad otherwise$

$$SSC = \sum_{n=2}^{N-1} f[(x_n - x_{n-1})x(x_n - x_{n+1})] \quad (7)$$

Where $f(x) = \{1 \quad , \quad if\ x \geq threshold\ 0 \quad otherwise$

$$Sk = \frac{\sum_{n=1}^{N} \left(x_n - \underline{x}\right)^{3\cdot} / N}{s^3} \quad (8)$$

Where $\underline{x}$ denotes the mean value and s represents the standard deviation of the sEMG signal.

$$Kr = \frac{\sum_{n=1}^{N} \left(x_n - \underline{x}\right)^{4} / N}{s^4} \quad (9)$$

$$IV = \sum_{n=1}^{N} |x_n| \quad (10)$$

2.3.3 Classification

Four classifiers were used to recognize the three words acquired from the two subjects. These classifiers were k-nearest neighbour (KNN), Neural network (NN), Support vector machine (SVM), and linear discriminate (LD). For each subject, 120 of the 150 records were used for training and 30 records were used for testing in a 5-Fold scheme.

All computations were done using MATLAB (Mathworks, Massachusetts, USA). The proposed system wasevaluated using several metrics, as follows:

Accuracy=	(TP+TN)/(TP+TN+FP+FN)	(11)
Sensitivity=	TP / (TP + FN)	(12)
Specificity =	TN / (TN + FP)	(13)
Precision=	TP / (TP + FP)	(14)
F1 Score =	2 TP / (2 TP + FP + FN)	(15)

Where TP: True positive, TN: True negative, FP: False positive, FN: False negative.

3. Results

Table 1: Results for Subject 1 / “Book” word

	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)
Ensemble	*86%*	80%	*89%*	*78.4%*	*79.2%*
NN	86%	84%	87%	76%	80%
SVM	81.3%	70%	87%	72.9%	71.4%
KNN	80.6%	70%	85%	70%	70%

Table 2: Results for Subject 1 / “Door” word

	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)
Ensemble	88%	83%	91%	82%	82%
NN	86.6%	81%	89%	78%	79.5%
SVM	86.6%	86%	87%	76%	81.1%
KNN	82%	76%	85%	71%	73.7%

Table 3 Results for Subject 1 / “Metro” word

	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)
Ensemble	86%	79.5%	*89%*	*78%*	*79.7%*
NN	*86%*	*80.8%*	*88%*	*76%*	*78.3%*
SVM	85.3%	*80%*	*87%*	*74%*	*77%*
KNN	86%	*79%*	*89%*	*78%*	*78.7%*

Table 4 : Confusion matrix of Ensemble classifier -Subject 1

		Predicted label		
		Book	Door	Metro
Actual label	Book	40	4	6
	Door	5	41	4
	Metro	6	5	39

Table 5 Confusion matrix of NN classifier -Subject 1

		Predicted label		
		Book	Door	Metro
Actual label	Book	42	4	4
	Door	6	39	5
	Metro	7	5	38

Table 6 Confusion matrix of SVM classifier -Subject 1

		Predicted label		
		Book	Door	Metro
Actual label	Book	**35**	**8**	**7**
	Door	**5**	**43**	**2**
	Metro	**8**	**5**	**37**

Table 7 Confusion matrix of KNN classifier -Subject 1

		Predicted label		
		Book	Door	Metro
Actual label	Book	**35**	**9**	**6**
	Door	**8**	**38**	**4**
	Metro	**4**	**7**	**39**

Table 8 Results for Subject 2 / “Book” word

	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)
KNN	*79.3%*	71.11%	*82.8%*	*64%*	*67.37%*
NN	*78.6%*	*68%*	*84%*	*68%*	*68%*
SVM	*78.6%*	*66%*	*85%*	*72%*	*69%*

Table 9 Results for Subject 2 / “Waseem” word

	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)
KNN	*71.33%*	66%	*74%*	*55%*	*60.5%*
NN	*76.67%*	*64%*	*83.8%*	*65%*	*64.6%*
SVM	*77.3%*	*70%*	*81%*	*64.8%*	*67.3%*

Table 10 Results for Subject 2 / “Metro” word

	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)
KNN	*73.33%*	60.8%	*78.85%*	*56%*	*58.3%*
NN	*74%*	*60.7%*	*80.8%*	*62%*	*61.39%*
SVM	*78.6%*	*66.67%*	*80.37%*	*57.14%*	*61.54%*

Table 11 Confusion matrix of Ensemble classifier -Subject 2

		Predicted label		
		Book	Waseem	Metro
Actual label	Book	**36**	**10**	**4**
	Waseem	**8**	**26**	**16**
	Metro	**8**	**8**	**34**

Table 12 Confusion matrix of KNN classifier -Subject 2

		Predicted label		
		Book	Waseem	Metro
Actual label	Book	**32**	**11**	**7**
	Waseem	**6**	**33**	**11**
	Metro	**7**	**15**	**28**

Table 13 Confusion matrix of NN classifier -Subject 2

		Predicted label		
		Book	Waseem	Metro
Actual label	Book	**34**	**8**	**8**
	Waseem	**6**	**32**	**12**
	Metro	**10**	**9**	**31**

Table 14 Confusion matrix of SVM classifier -Subject 2

		Predicted label		
		Book	Waseem	Metro
Actual label	Book	**36**	**6**	**8**
	Waseem	**9**	**35**	**6**
	Metro	**9**	**13**	**28**

4. Discussion and Conclusion

This preliminary study employed two speech impaired subjects whose attempts to generate spoken words produce sounds that are mostly unintelligible by a typical listener. The goal of the study was to show that using standard machine learning techniques it could be possible to recognize words uttered by the two subjects in a limited vocabulary dataset. Rather than using a microphone to obtain audio signals corresponding to the uttered words, we acquired surface electromyography signals generated by muscles surrounding the vocal cords upon the utterance of the words. This can help prevent unavoidable audio interferences from background sources, which can greatly limit the utility of the proposed system. We attempted word-level recognition using standard machine learning techniques. The intended solution is meant to be adapted for each speech impaired user, as their attempts to generate word utterances vary in the degree of success for different words.

It should be noted that the effort exerted by a speech impaired person when trying to generate word utterances is quite substantial compared to non-speech impaired persons. While this may not be a problem in a typical scenario, the case of acquiring data for training and testing is challenging as they are asked to repeat the intended words 50 times each in the same session. This limits the size of the data that can be obtained and requires a good amount of motivation to keep the subjects engaged. This fact, however, does not eliminate that fact that one of the main weaknesses of this study is the limited size of the used dataset. Nevertheless, the small number of subjects and the limited number of words used still show that the hypothesis that a well trained engine should be able to accurately recognize word utterances (with some system personalization) could be feasible. While the accuracy shown in the results of this paper are not high by typical speech recognition methods for the non-imaired persons (now close to 100% in many settings), they are considered very promising given the nature of the problem and the lack of prior research guidance. These promising results motivate future work using larger datasets and more speech impaired subjects. They also motivate building analysis frameworks that better suit the nature of the speech and they type of the acquisition sensor.

Acknowledgements

This research was made possible by a research grant from the Ministry of Communication and Information Technology under the ITAC PRP funding program, grant code CFP190-Myo-Speech.

References

[1] Shakeel Sheikh, Md Sahidullah, Fabrice Hirsch, Slim Ouni. Machine Learning for Stuttering Identification: Review, Challenges & Future Directions. 2022. hal-03634072f

[2] World Health Organization, 2021, WHO, (1 April 2021), [Online]. Available on: <https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss>

[3] McLean, S. *The basics of speech communication*. Boston, MA: Allyn & Bacon, (2003).

[4] Ceil Lucas, *The Sociolinguistics of the Deaf Community*, (1995) p. 80.

[5] Pearson, J., & Nelson, P. *An introduction to human communication: Understanding and sharing* . Boston, MA: McGraw-Hill (2000) p. 6.
[6] Meltzner, G.S., Sroka, J., Heaton, J.T., Gilmore, L.D., Colby, G., Roy, S., Chen, N., Luca, C.J.D. "Speech recognition for vocalized and subvocal modes of production using surface EMG signals from the neck and face", In *INTERSPEECH*, pp.2667-2670,2008.
[7] M. S. Elmahdy and A. A. Morsy, "Subvocal speech recognition via close-talk microphone and surface electromyogram using deep learning," in *Federated Conference on Computer Science and Information Systems* (FedCSIS), 2017, pp. 165-168
[8] M. Wand, C. Schulte, M. Janke, and T. Schultz, "Array-based Electromyographic Silent Speech Interface," in *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing*, 2013, pp. 89–96.
[9] Dezhen Xiong, Daohui Zhang, Xingang Zhao, Yiwen Zhao, "Deep Learning for EMG-based Human-Machine Interaction: A Review", in *IEEE/CAA Journal of Automatica Sinica*, vol.8, no.3, pp.512-533, 2021.
[10] Pasinetti, S.; Lancini, M.; Bodini, I.; Docchio, F. A Novel Algorithm for EMG Signal Processing and Muscle Timing Measurement. *IEEE Trans.* Instrum. Meas. 2015, 64, 2995–3004.
[11] Shair, E.; Ahmad, S.; Abdullah, A.; Marhaban, M.; Tamrin, S.M. "Determining Best Window Size for an Improved Gabor Transform in EMG Signal Analysis", in *TELKOMNIKA Telecommun. Comput. Electron. Control*. 2018, 16, 1650.
[12] A. Phinyomark, C. Limsakul, and P. Phukpattaranont, "A Novel Feature Extraction for Robust EMG Pattern Recognition", in *Journal of computing*, vol. 1, no. 1, 2009.