2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

A Pilot Study on the Performance of Time-Domain Features in Speech Recognition based on high-density sEMG

Xiaochen Wang[#], Mingxing Zhu[#], Oluwarotimi Williams Samuel, Zijian Yang, Lin Lu, Xingxing Cai, Xin Wang, Shixiong Chen[*], *IEEE Member*, Guanglin Li, *IEEE Senior Member*

***Abstract*— Features extracted from the surface electromyography (sEMG) signals during the speaking tasks play an essential role in sEMG based speech recognition. However, currently there are no general rules on the optimal choice of sEMG features to achieve satisfactory performance. In this study, a total of 120 electrodes were placed on the face and neck muscles to record the high-density (HD) sEMG signals when subjects spoke ten digits in English. Then ten different time-domain features were computed from the HD sEMG signals and the classification performance of the speech recognition was thoroughly compared. The contribution of each feature was examined by using three performance metrics, which include classification accuracy, sensitivity, and F1-Score. The results showed that, among all the ten different features, the features of WFL, MAV, RMS, and LOGD were considered to be superior because they achieved higher classification accuracies with high sensitivities and higher F1-Scores across subjects/trials in the sEMG-based digit recognition tasks. The findings of this study might be of great value to choose proper signal features that are fed into the classifier in sEMG-based speech recognition.**

***Clinical Relevance* — This pilot study proved that WFL, MAV, RMS, and LOGD might be the optimal features to extract from sEMG signals for sEMG-based speech recognition to achieve satisfactory performance in different applications.**

I. Introduction

In recent years, the development of speech recognition technology based on surface electromyography (sEMG) signals has been progressing rapidly [1, 2]. This technique requires extracting effective feature information from the sEMG signal and inputting it into the classifier for classification and recognition [3]. Therefore, feature extraction and classifier selection are two essential links, which have a great impact on the final recognition effect [4]. Feature extraction is an important method to extract useful information hidden in the sEMG signal [5]. Generally, features in the analysis of the sEMG signal are plentiful, and there are time-domain, frequency-domain, time-frequency, and time-scale [6]. In order to classify the sEMG signals successfully, the selection of the features is necessitated to be concerned prudently. However, many studies on the classification of sEMG signals have used feature sets containing a large number of redundant features, resulting in poor classification or excessive calculation [7, 8]. Therefore, it is necessary to study the effects of the features on speech recognition to keep away from using the bad features in the classification stage.

In the field of speech recognition using sEMG signals, many efforts had been made to improve the classification effect by selecting multiple features, while there had few in-depth studies which make quantitative comparisons of the effects of feature selection on recognition qualities. Soon et al. extracted six time-domain features from sEMG signals for classifying Malay words, but the features had inputted the classifier as a whole dataset and the contribution of individual feature to the classification result were ignored [9]. Srisuwan et al. investigated the performance of eight features using sEMG signals for classifying the Thai tonal sound [10]. However, only four electrodes were stuck on the neck of the subjects, and it had not been verified that the extraction of these features would have similar effects at different locations. Thus, further exploration was obliged to make up for the above shortcomings.

In this paper, the High-Density (HD) sEMG signals were recorded by a total of 120 surface electrodes located on the facial and neck muscles when the subjects spoke ten digits in English. Then, ten time-domain features that were generally analyzed in sEMG signals were extracted to input into the LDA classifier to classify ten speaking tasks. Finally, the effectiveness of each feature was verified by using three performance metrics, they were classification accuracy, sensitivity, and F1-Score namely.

II. Methods

A. Signal Measurement

19 healthy subjects (twelve males, seven females) aged from 22 to 26, with a mean age of 25.25, were recruited for the

* This work was supported in part by the National Natural Science Foundation of China (#61771462, #61901464, #U1613222), Shenzhen Governmental Basic Research Grant (#JCYJ20180507182241622), Science and Technology Planning Project of Shenzhen (#GJHZ20190821160003734), Science and Technology Program of Guangzhou (#201803010093), Science and Technology Planning Project of Guangdong Province (#2019A050510033), SIAT Innovation Program for Excellent Young Researchers (E1G027).

X. C. Wang, M. Zhu, O. W. Samuel, Z. Yang, X. Wang, S. Chen, and G. Li are with the CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China (Dr. Shixiong Chen, phone: 86-18566773423; e-mail: sx.chen@siat.ac.cn).

X. C. Wang, M. Zhu, and X. Wang are also with Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China.

X. C. Wang, M. Zhu, O. W. Samuel, Z. Yang, X. Wang, S. Chen, and G. Li are with the Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen, Guangdong 518055, China.

L. Lu and X. Cai are with Shenzhen Nanshan People' Hospital, Huazhong University of Science and Technology Union Shenzhen Hospital, Shenzhen, Guangdong 518052, China.

# The first two authors contributed equally to the work.

experiments. All the subjects were kept in a healthy state and had no pronunciation difficulties. Before the experiment, all participants were fully informed of the experimental procedures, and they voluntarily signed the informed consent forms to allow the publishing of their data for research purposes. The protocol of this experiment was approved by the Institutional Review Board of Shenzhen Institutes of Advanced Technology, Chinese Academy of Science.

B. Experimental Procedure

The subjects were covered by 120 channels of electrodes from a multi-channel data acquisition system (TMSi, REFA, the Netherlands), which consists of two 4*5 matrices on the face region and two 5*8 arrays on the muscles in the front neck area (as shown in Fig. 1). The whole distribution of electrodes was almost 2-D with the horizontal and vertical distance between two adjacent electrodes both approximately 15mm, and electrodes arrays were separated into three types based on their locations, including the face-group (40 channels), neck-group (80 channels), and whole-group (120 channels). The HD sEMG data was acquired from the high-density electrodes with a sampling rate of 2048 Hz for each channel, and all possible articulatory muscles on both sides of the face and neck were covered to ensure the completeness and fullness of data. A fabric electrode was attached on the left wrist of the subject as a reference to collect data.

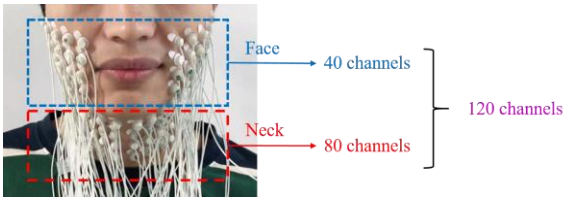

Figure 1. The positions of the electrodes in tasks

Before the experiment, the subjects' skin was prepared by wiping using an alcohol pad to minimize skin-to-electrode impedance between electrodes and skin, and they were required to sit on a chair quietly and maintain muscle relaxation. During the procedure, the sEMG signals in a silent mode were recorded with no speaking and body moving for almost 40 seconds, which is treated as the baseline. The subjects were required to complete ten sets of English pronunciation tasks, corresponding to the pronunciation of English digits from zero to nine. Each set of pronunciation tasks includes a repetition of 28 times, and each one-second speaking was followed by a three-second break. The HD sEMG signals were obtained by the group of electrodes distributed all over the whole face and neck when each recruited subject was doing speaking tasks in a normal volume.

C. Signal processing and data analysis

The original HD sEMG signals were preprocessed by two processes. First, a fourth-order bandpass Butterworth filter with a cut-off frequency of 50-500HZ was utilized, which can filter out low-frequency noise below 50HZ, such as interference from ECG and the artifacts. Then, the notch filter of 50HZ and it's integer multiples were arranged to filter out the power line interference and it's harmonics. After the preprocess, cleaner signals were obtained. The filtered signals of 28 repetitions were manually sliced for each digit with only the speaking parts reserved and recombine them. Subsequently, the sEMG features were calculated using a series of sliding windows segmented from the processed data with a length of 400 sampling points and an overlap-interval of 200 sampling points. The calculation of 10 time-domain features that contained the valuable information for speech recognition was based on the analysis windows. The 10 time-domain features and their mathematical definition were as shown in Table I.

TABLE I. TEN TIME-DOMAIN FEATURES WITH THEIR MATHEMATICAL DEFINITIONS [11]

S/N	Time-domain feature	Abbr	Mathematical expression
F1	Waveform Length	WFL	$\mathrm{WFL}=\sum_{n=1}^{k-1}\left[f\left(\lvert x_{n+1}-x_n\rvert\right)\right]$; $f(x)=\begin{cases}1, & \text{if } x\geq Thr.\\ 0, & \text{otherwise}\end{cases}$
F2	Mean Absolute Value	MAV	$\mathrm{MAV}=\frac{1}{k}\sum_{n=1}^{k}\lvert x_n\rvert$
F3	Root mean square	RMS	$\mathrm{RMS}=\sqrt{\frac{1}{k}\sum_{n=1}^{k}x_n^2}$
F4	Logarithm Detector	LOGD	$LOGD=e^{\frac{1}{k}\sum_{n=1}^{k}\log(\lvert x_n\rvert)}$
F5	Simple Square Integral	SSI	$\mathrm{SSI}=\sum_{n=1}^{k}x_n^2$
F6	Variance	VAR	$\mathrm{VAR}=\frac{1}{k-1}\sum_{n=1}^{k}x_n^2$
F7	Slope Sign Changes	SSC	$\mathrm{SSC}=\sum_{n=2}^{k-1}\left[f\left(x_n-x_{n-1}\right)^{*}\left(x_n-x_{n+1}\right)\right]$
F8	Zero Crossings	ZCS	$\mathrm{ZCS}=\sum_{n=1}^{k-1}\left[\mathrm{sgn}\left(x_n*x_{n+1}\right)\cap\left(x_n-x_{n+1}\right)\geq \mathrm{Thr.}\right]$
F9	Third Moment	TM3	$\mathrm{TM3}=\left\lvert\frac{1}{k}\sum_{n=1}^{k}x_n^{3}\right\rvert$
F10	Kurtosis	KURT	$KURT=\sum_{n=1}^{k}\frac{E\left(x_n-\mu\right)^4}{\sigma^4}$

Afterward, 5-fold cross-validation was employed in this analysis to reduce the variability of the data and avoid over-fitting. These datasets were subsequently input into the linear discriminant analysis (LDA) classifier for recognizing the speech. Finally, to determine the effectiveness of each feature, three performance metrics, including classification accuracy (CA), sensitivity (Sen), and F1-Score (F1score) were applied, and they were mathematically expressed as in Table Ⅱ. tp represented the number of correct predictions per class in the multi-class confusion matrix, and α denoted the corresponding error values.

TABLE II. THREE PERFORMANCE METRICS WITH THEIR MATHEMATICAL DEFINITIONS [12]

S/N	Mathematical expression
1	$CA=\frac{\text{Number of correctly classified samples}}{\text{Total number of testing samples}}100\%$
2	$\mathrm{Sen}=tp_1/\left(tp_1+\alpha_{12}+\alpha_{13}+\alpha_{14}+\alpha_{15}\right)$
3	$F1_{\mathrm{Score}}=2*\frac{\mathrm{Precision}\ *\ \mathrm{Recall}}{\mathrm{Precision}\ +\ \mathrm{Recall}}$
4	$\mathrm{Precision}\ =tp_1/\left(tp_1+fp_1\right)$
5	$fp_1=\alpha_{21}+\alpha_{31}+\alpha_{41}+\alpha_{51}$
6	$\mathrm{Recall}=\mathrm{Sen}$

III. Results

A. The comparison of classification accuracies across ten features in recognizing ten English words

In this part, the classification accuracies of ten words using ten features across one subject based on sEMG signals recorded from different regions were studied, as shown in Fig. 2. The dots represented the classification accuracies of ten words. The magenta dots, red dots, and blue dots signed the CAs calculated by using sEMG signals collected from whole-group, neck-group, and face-group, respectively. For the given speaking task, the CAs from the whole-group could be found to have better performance than the neck-group and face-group. Besides, it was observed that the CAs showed a decreasing trend from left to right (F1 to F10). Moreover, the CA of F1 ranked number one with average accuracies higher than 85%. Whereas F9 and F10 achieved apparently lower CAs than the other eight features. Meanwhile, there were more discrete dots of the CAs of F7, F8, F9, and F10. All the mentioned circumstances could be clearly observed in all three groups.

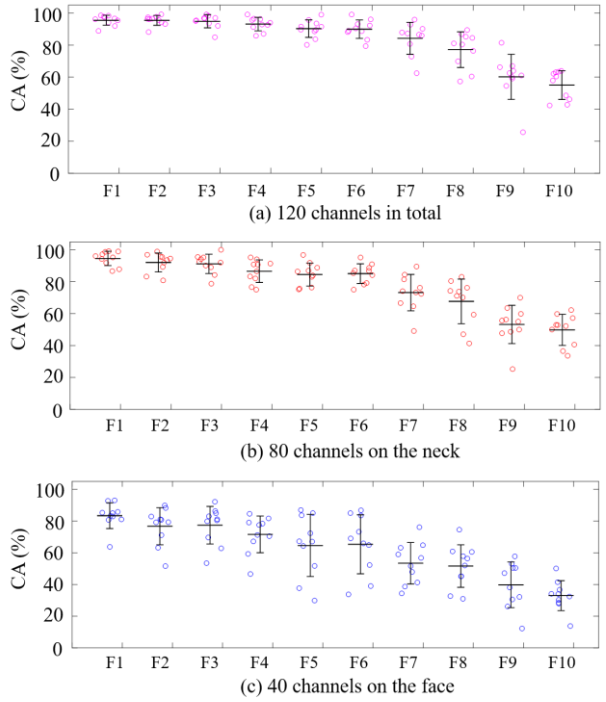

Figure 2. The comparison of classification accuracies of ten words using ten features across one subject with electrodes on different groups: (a) 120 channels in total, (b) 80 channels on the neck, (c) 40 channels on the face

B. Comparing the distributions of classification accuracies of nineteen subjects using ten features

The average classification accuracies of ten words using ten features across 19 subjects with electrodes on different groups were analyzed in this section for further exploring the influence of features on classification performance during speech recognition. The statistical distribution of the average CAs of all the ten speaking tasks across 19 subjects of 10 features was shown in Fig. 3. The boxplot consists of five numerical positions, namely the minimum observation (lower edge), 25% quantile (Q1), median, 75% quantile (Q3), and maximum observation (upper edge). It was observed that the average CAs from the whole-group were higher than the neck-group and face-group across all the features. Moreover, the values of CA gradually decreased from left to right (from F1 to F10). The CAs of F1 was the highest among the ten features, no matter for the whole-group, neck-group, or face-group. A similar circumstance could also be found in Fig. 3(b) and Fig. 3(c).

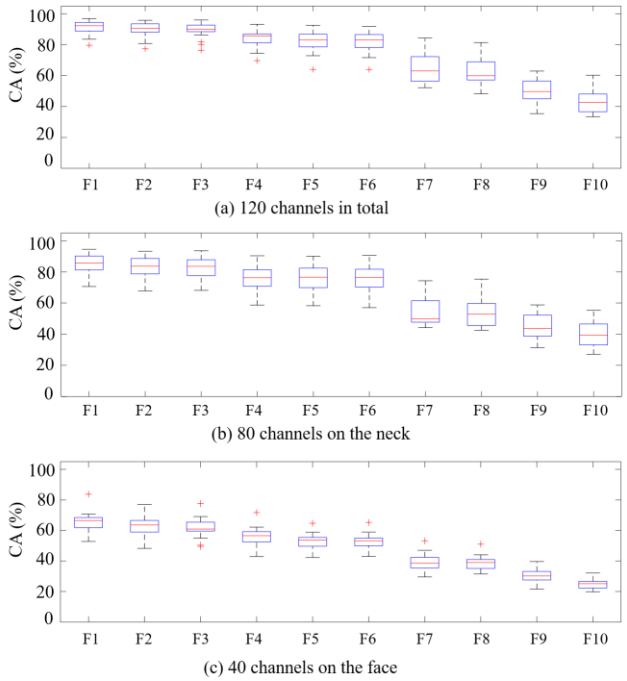

Figure 3. The distributions of classification accuracies of nineteen subjects using ten features with electrodes on different groups: (a) 120 channels in total, (b) 80 channels on the neck, (c) 40 channels on the face

C. Classification performance of the ten features based on Sensitivity and F1-Score metrics

To further analyze the performance of features, the ten individual features were assessed with two additional metrics, namely the sensitivity and F1-Score, and the obtained results are shown in Fig. 4. The results showed that the highest values of sensitivity and F1-Score both appeared in F1. Besides, the values of sensitivity and F1-Score in F2, F3, and F4 were slightly smaller than in F1, which is as same as what could be noticed in the aforementioned CAs. It was noteworthy that the standard deviations of sensitivity and F1-Score achieved the smallest values in F1, and the values in F2, F3, and F4 were a little higher than in F1. Moreover, the sensitivity and F1-Score obtained for individual features across 19 subjects were ranked in decreasing order from F1 to F10 according to Fig. 4. Both the values of sensitivity and F1-Score showed bad performance in F7, F8, F9, and F10.

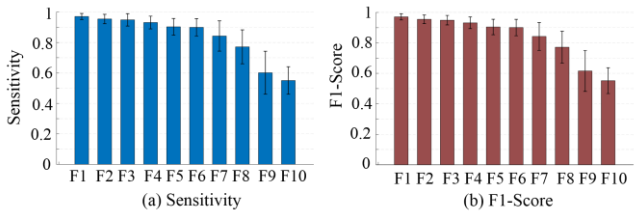

Figure 4. Classification performance of the ten features based on Sensitivity and F1-Score metrics: (a) Sensitivity results; (b) F1-Score results

IV. DISCUSSION

Features extracting from sEMG signals could deliver useful information hidden in the muscle activities [13]. Thus, the features extracted from the sEMG signals during the speaking tasks play an important role in speech recognition, which even have an enormous impact on the performance of recognition. In order to provide a reference for selecting features for subsequent research on using EMG signals for speech recognition, the performance of 10 time-domain features commonly utilized in the classification of speech recognition was compared in this pilot study.

In this study, the classification accuracies of ten words using ten features across one subject based on sEMG signals recorded from different regions were first compared. Results showed that the CAs of F1 (WFL), F2 (MAV), F3 (RMS), and F4 (LOGD) always had better performance than the other six features with electrodes on different regions. The result was consistent with the fact that WFL, MAV, and RMS were suitable for identifying the flexion and extension of the elbow than ZC proved in Castro's research [14]. For the sEMG signals collected at different electrode positions, the CAs obtained from the features extracted from them had similar characteristics. This outcome showed that the results in this article could be extended to different electrode positions, thereby avoiding the harm caused by ignoring electrode position changes that might affect the feature performance in Srisuwan's study [10]. Besides, there was a similar trend on the average CAs across 19 subjects, which indicated that the relative magnitudes of the classification accuracy values when different subjects used these ten features for recognition were similar and not affected by individual differences as a whole. In order to evaluate the properties of the ten features in sensitivity and F1-Score, the histogram was represented. The results told the truth that a similar trend with the sensitivities and F1-Score where F1, F2, F3, and F4 gave better performance, just as same as what could be noticed in the aforementioned CAs. It is generally accepted that a single feature is outstanding when it achieved high classification accuracies with high sensitivity and high F1-Score across subjects/trials after been subjected to different testing conditions [12]. For the evidence provided above, the conclusion could be drawn that F1, F2, F3, and F4 might be the most suitable features to extract from surface EMG and input it into LDA classifier for speech recognition among the ten features compared in this paper.

Results of this study can be widely used and applied in many sEMG signal classification studies, including medical and engineering applications, to keep away from using the bad features in the classification stage. However, features in the analysis of the sEMG signal are generally plentiful, but only ten time-domain features were considered in this pilot study, which might be the limitation. The limitation will be further studied in our future research.

V. CONCLUSION

In this study, the performance of 10 time-domain features commonly utilized in classification in speech recognition was compared. The results confirmed that WFL, MAV, RMS, and LOGD might be the most appropriate features extracted from sEMG signals and input into the LDA classifier for speech recognition among the ten time-domain features compared in this paper. The conclusion might be of great benefit to select proper features in speech recognition based on sEMG signals.

REFERENCES

[1] G. S. Meltzner, J. T. Heaton, Y. Deng, D. L. Gianluca, S. H. Roy, and J. C. Kline, "Development of sEMG sensors and algorithms for silent speech recognition," *Journal of Neural Engineering*, vol. 15, no. 4, pp. 046031.1-046031.11, 2018.

[2] H. Manabe, A. Hiraiwa, T. Sugimura, "Unvoiced speech recognition using EMG - Mime Speech Recognition," in Extended Abstracts of the Conference on Human Factors in Computing Systems, 2003, pp. 794-796.

[3] N. S. Jong, P. J. B. Phukpattaranont, and B. Engineering, "A speech recognition system based on electromyography for the rehabilitation of dysarthric patients: A Thai syllable study," *Biocybernetics and Biomedical Engineering*, vol. 39, no.1, pp. 234-245, 2018.

[4] R. Boostani and M. H. Moradi, "Evaluation of the forearm EMG signal features for the control of a prosthetic hand," *Physiological Measurement*, vol. 24, no. 2, pp. 309-319, 2003.

[5] M. Zardoshti-Kermani, B. C. Wheeler, K. Badie, and R. M. Hashemi, "EMG feature evaluation for movement control of upper extremity prostheses," *IEEE Transactions on Rehabilitation Engineering*, vol. 3, no.4, pp. 324-333,1995.

[6] M. A. Oskoei and H. Hu, "GA-based Feature Subset Selection for Myoelectric Classification," in IEEE International Conference on Robotics & Biomimetics, 2006, pp. 1465-1470.

[7] Y. C. Du, C. H. Lin, L. Y. Shyu, and T. J. E. S. w. A. A. I. J. Chen, "Portable hand motion classifier for multi-channel surface electromyography recognition using grey relational analysis," *Expert Systems with Applications,* vol. 37, pp. 4283-4291, 2010.

[8] T. Kubo, M. Yoshida, T. Hattori, and K. Ikeda, "Feature Selection for Vowel Recognition Based on Surface Electromyography Derived with Multichannel Electrode Grid," in International Conference on Intelligent Science and Intelligent Data Engineering, 2011, pp. 242-249..

[9] M. W. Soon, M. I. H. Anuar, M. H. Z. Abidin, A. S. Azaman, and N. M. Noor, "Speech recognition using facial sEMG," in 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 2017, pp. 201-205.

[10] N. Srisuwan, P. Phukpattaranont, and C. J. P. E. Limsakul, "Feature selection for Thai tone classification based on surface EMG," *Procedia Engineering*, vol. 32, pp. 253-259, 2012.

[11] A. Phinyomark, P. Phukpattaranont, and C. J. E. s. w. a. Limsakul, "Feature reduction and selection for EMG signal classification," *Expert systems with applications*, vol. 39, no. 8, pp. 7420-7431, 2012.

[12] O. W. Samuel, Y. Geng, X. Li, and G. J. J. o. m. s. Li, "Towards efficient decoding of multiple classes of motor imagery limb movements based on EEG spectral and time domain descriptors," *Journal of medical systems*, vol. 41, no. 12, pp. 1-13, 2017.

[13] Z. Ding, C. Yang, Z. Tian, C. Yi, Y. Fu, and F. J. S. Jiang, "sEMG-based gesture recognition with convolution neural networks," *Sustainability*, vol. 10, no. 6, p. 1865, 2018.

[14] M. C. F. Castro, E. L. Colombini, P. T. Junior, S. P. Arjunan, and D. K. J. B. E. O. Kumar, "sEMG feature evaluation for identification of elbow angle resolution in graded arm movement," *BioMedical Engineering OnLine*, vol. 13, pp.155-165, 2010.