Attention Bidirectional LSTM Networks Based Mime Speech Recognition Using sEMG Data

Hongyi Ye[1], Haohong Lin[1], Zijun Song[1], Ming Zhang[1], Ruifen Hu[1], Nan Li[2] and Guang Li[1]
[1]*Institute of Cyber-system and Control, Zhejiang University, Hangzhou, China*
[2]*Department of Engineering, University of Cambridge, Cambridge, UK*
Corresponding Email: guangli@zju.edu.cn

Abstract—**Surface electromyography (sEMG) has been proven competent and reliable to recognize speech musculature movement patterns. In other words, we can understand what a person prepares to say by collecting sEMG signals around the mouth. Therefore, sEMG-based Mime Speech Recognition (MSR) is a potential technique for human-machine interaction within noisy surroundings as well as the application of helping dysarthric patients. In this paper, we introduce multi-layer Bidirectional Long Short-Term Memory (BLSTM) networks with attention mechanism as a classifier for MSR, and verify it in the data set collected by ourselves. Six-channel sEMG signals are firstly acquired from elaborately selected facial muscles. Short-time Fourier Transform (STFT) and Convolutional Neural Networks (CNN) are utilized to extract time-frequency domain feature maps, replacing the handcrafted features in classic methods. The second phase of recognition process lies in the designed classifier. This classification system achieves over 97% accuracy in the four-class MSR task, significantly surpassing simple CNN and LSTM methods. Such result also indicates that excellent MSR results can be achieved without relying on handcrafted signal features.**

Index Terms—**surface electromyography, mime speech recognition, human-machine interaction, attention BLSTM**

I. Introduction

Speech recognition technology has been gradually applied to our life since its development in the late 20th century. However, there have been many difficulties for speech recognition in the transition from laboratory to applications in daily lives. One of the most important factors is the existence of various sorts of noise. Affected by other unrelated sounds, the performance of speech recognition systems based on acoustic signal patterns will be greatly reduced [1].

To tackle this problem, one idea is to use bioelectric signal sensors to obtain what people want to say through Mime Speech Recognition (MSR). MSR method is advantageous since it utilizes information from the nervous system of muscle endplate instead of the acoustic system, hence neglecting the environmental noise. Moreover, it offers a way to communicate privately without disturbing bystanders and provide voice communication for people with severe speech impairments (e.g., laryngectomy patients) [2]. Compared with normal speech recognition, the disadvantage of MSR is that it cannot recognize many words. With the increase of vocabulary, the recognition accuracy decreases rapidly. Therefore, MSR is more suitable to identify specific commands, especially in control systems and robots.

Electromyography (EMG) signals, considered as a secondary source of speech information, are usually used for MSR [3]. In general, there are two main methods to obtain EMG signals. The first one is invasive. During the test, electrodes are inserted through the skin and go into the muscle. Though it can collect signals accurately, it will easily lead to injury. Therefore, it is unsuitable for ordinary human testers. The other one is non-invasive. This method captures surface electromyography (sEMG) signals. Without injury and complicated electrode placement process, the sEMG is popular to users and researchers. The sEMG sensors placed on subjects' vocalization-related muscles are designed to detect the output of their sEMG signals when subjects speak different words silently. Then we solely use the collected signals to conduct MSR tasks.

In this study, we mainly focus on the feature extraction in time-frequency domain data, generated by a Short-time Fourier Transform (STFT). By using feature extraction on filtered time series for specific channels, we don't need to make complex conversions on raw data. However, we still get desired information on specific channels and build up the robustness of signal recognition process.

The researchers have done several studies on sEMG-based speech recognition for isolated words and phrases. As early as 1991, Morse used artificial neural network (ANN) to recognize ten English words represented by sEMG with 60% accuracy [4]. Later, Kumar and his colleagues solely utilized sEMG signals to recognize five English vowels. They used root mean square (RMS) values as features and ANN as the classifier, achieving 88% accuracy [5]. In the past 20 years of research, hidden Markov model (HMM) was most often used by researchers for sEMG-based speech recognition [6] [7] [8]. Based on HMM, Lee expanded the number of MSR of isolated Korean words to sixty, gaining 87% accuracy [9]. What these studies had in common was that they needed to select handcrafted signal features carefully, such as RMS, zero crossings (ZC), and even mel frequency cepstral coefficient (MFCC). In recent years, as deep learning methods prevailing, Ai tried to utilize CNN to do MSR [10], while Cao employed LSTM for sEMG recognition [11]. In special cases, Jong indicated that sEMG-based MSR can achieve similar results in normal people and people with dysarthric disturbances [12].

This work is supported by the Science Fund for Creative Research Group of NSFC (Grant No. 61621002) and Zhejiang University Education Foundation Global Partnership Fund.

In addition, Meltzner proved sEMG-based MSR could work on laryngectomy patients [13].

The conventional framework of sEMG-based recognition system is usually composed of signal detection, signal preprocessing, segmentation, feature extraction and classification. The features used for classification are generally hand-crafted by human experts [14]. And it took a lot of time and efforts to select proper features. In order to obtain more insightful and efficient features in complex datasets, deep learning methods are introduced to carry out sEMG-based MSR. We introduce the Bidirectional LSTM (BLSTM) together with self-attention layers to achieve a better performance on the classification accuracy on test dataset.

In summary, our contribution lies in the following aspects:

- We apply multi-layer attention BLSTM networks as the classifier in sEMG-based MSR, obtaining higher classification accuracy than CNN, LSTM and Random Forest.
- We demonstrate that excellent MSR results can be achieved without relying on manual signal features as former studies did, saving time and energy.

II. Data Collection and Preprocessing

A. Corpus and data set

In this study, we choose four Chinese words for isolated vocabularies recognition task. The long-term goal is to use these words for robot arm control. Also, we select words that most people can meet the standard Mandarin pronunciation, because users from different places can pronounce the same word very differently. “TABLE I” demonstrates the information of the corpus.

TABLE I
Vocabulary List

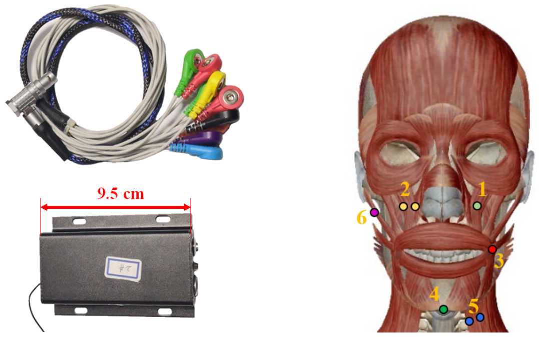
Type No.	Chinese Pinyin	Pronunciation	Meaning
1	Er	/ɛr/	Two
2	Yi	/i:/	One
3	Kuai	/kuaI/	Fast
4	Man	/man/	Slow

The words “Yi” and “Er” are intended to determine the serial number of robot arms controlled, while “Kuai” and “Man” are going to set their move speed. As for the subjects, 7 testers contain 3 males and 4 females, with an average age of 22. The data set includes 6000 samples totally, with 1500 samples in each category. The length of each sample is 2000 milliseconds, with the sampling rate in 1000Hz.

B. sEMG data acquisition

The collection device consists of circuit modules, removable wires and non-invasive Ag/AgCl sensors in our experiments [15]. During MSR experiments, subjects cyclically and speak four words in the corpus silently at the prompt. Once they start mime speech, researchers can obtain six channels of sEMG data from ten electrodes placed on the subjects’ facial muscles who related to phonation. The channel distribution shown in “Fig. 1(b)” has been proven useful to obtain high quality sEMG signals [16] [17] [18]. On channel no.2 and no.5, two sensors are placed to sample differential signals. There are also two floating ground reference electrodes sticking on ears’ mastoid portion. The Captured signals are transmitted to a computer through wireless fidelity communication, where a LabVIEW application visualizes sEMG signals and gives subjects the next instruction.

In the experiment, sedentary subjects are not allowed to make sound, and also try to minimize their muscle movements during mime speech to simulate real-world applications.



(a) Collection device

(b) Detection areas

Fig. 1. (a) shows the size of the portable collection device, whose diameter is within 10cm and weight is less than 500g. (b) illustrates electrode locations on surface of face and neck. The muscle names corresponding to each sampling channel are: 1,2- levator labii superioris and levator angularis; 3- the interaction of orbicular oris, zygomaticus minor, zygomaticus major, risorius, buccinator and depressor anguli oris; 4- mentalis; 5- platysma; 6- zygomaticus major and masseter pars superficialis.

C. Signal preprocessing

To avoid the influence of power frequency noise, ECG noise as well as noise by user’s body movement, we designed a Notch Filter at the frequency of 50Hz, followed by a Bandpass filter. The framework of preprocessing illustrated below resists most sorts of noise.

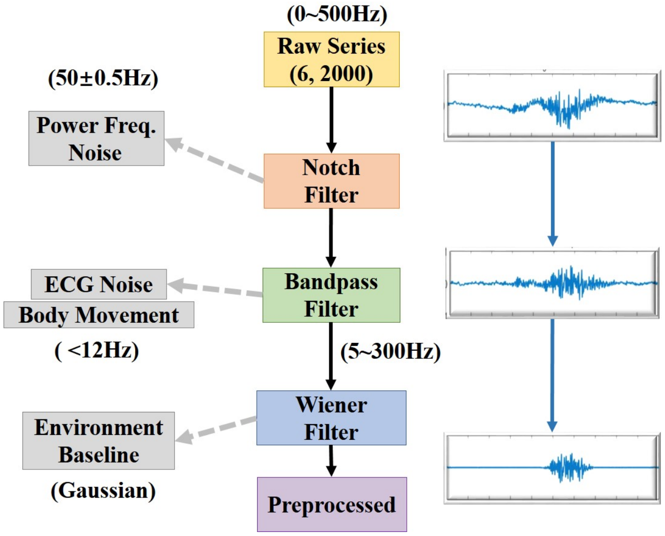


Fig. 2. The sEMG signal preprocessing steps.

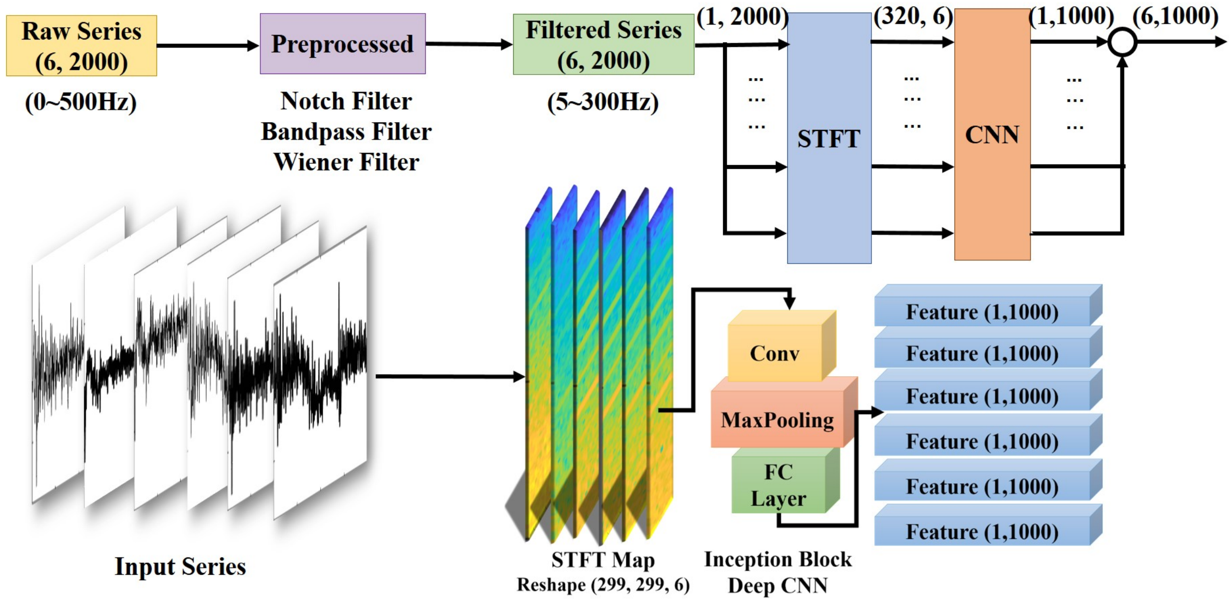


Fig. 3. The overview of preprocessing and the automatic feature extraction procedure. After preprocessing, we use STFT and CNN successively to extract time-frequency feature maps. The CNN is pre-trained, with fixed parameters. Every step's output size is shown in the figure.

We adopt a 21st order Butterworth Filter as our Bandpass Process, whose Square amplitude frequency characteristic could be described as:

$$|H(\omega)|^2 = \frac{1}{1+(\frac{\omega}{\omega_c})^n} \ , \ where\ n = 21 \qquad (1)$$

Another possible noise comes from baseline environmental noise. With the technique of wiener filter, we could reduce the noise of raw time series for each channel. Based on the principle of posterior noise spectral subtraction as well as the minimum mean square error, we could get satisfied de-noised results that could serve well for further feature extraction engineering. Here we select the window length of wiener filter as 256ms and model the default noise as additive white Gaussian noise with standard deviation as 1. We could get an MMSE result under a proper setting of Wiener Filter.

III. Speech Recognition Model

The MSR model can be divided into two parts: automated feature extraction and classification. The STFT and CNN with fixed parameters can extract insightful time-frequency domain features automatically [19]. Then, four-layer BLSTM networks with self-attention mechanism serve as a classifier to sort sEMG data. The whole process is illustrated in "Fig. 3" and "Fig. 4".

A. Feature extraction

Short Time Fourier Transform, usually known as STFT, is a common tool that segments time series into several shorter sections, each of which could indicate the frequent patterns varying from time. The STFT is defined as:

$$X_m(\widehat{\omega}) = \sum_{n=-\infty}^{\infty} x(n) w_M(n-mR) e^{-j\omega n} \qquad (2)$$

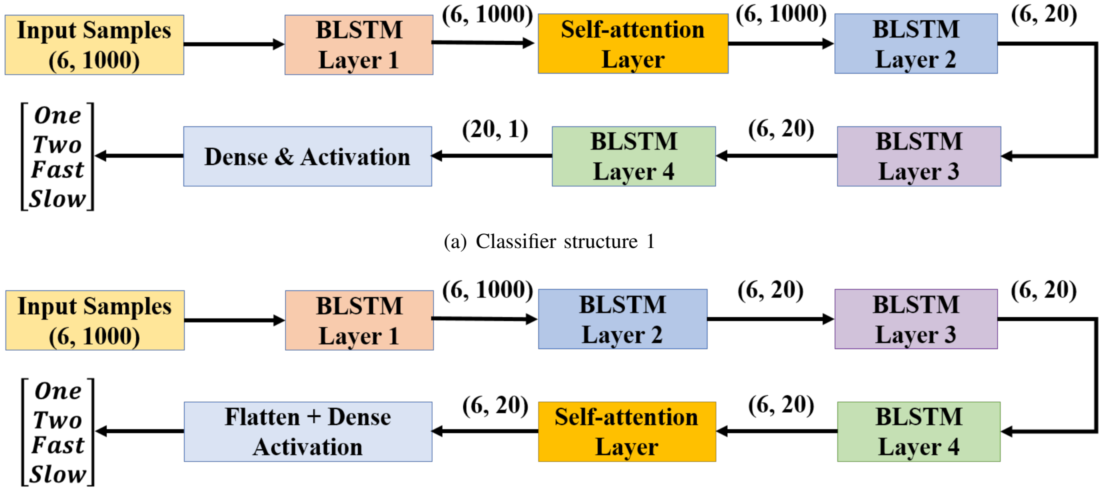
Where $x(n)$ stands for input time series; $w_M(n)$ means the window function like Blackman with a length of M; R represents the hop size between successive STFT; and $X_w(\widehat{\omega})$ indicates the DTFT windowed data centered at time $t_c = mR$.

We find a stabilized time series when we set window size $\mathrm{M} = 256\mathrm{ms}$ and hop size $\mathrm{R} = 192\mathrm{ms}$. The shape of $X_m(\widehat{\omega})$ could be calculated as: $\left(1+\frac{M+R}{2}, \left[\frac{\mathrm{N}}{R}\right]+1\right) = (320, 11)$. Due to the symmetricity of FFT, we could then analyze the first six frames, that is, a matrix of $(320, 6)$. For each channel of input data, we carry out the STFT to transform the original data to a time-frequency domain matrix, containing some information of time and frequency property for independent axis. For such a time-frequency domain matrix, we further extract features with CNN [20], which extracts features in time-frequency domain simultaneously.

By doing a channel-specific feature extraction, we concatenate the results together and get the output data in the feature space. The final size of the output feature map is $(6, 1000)$ each sample, as is illustrated in "Fig. 3". These features could further serve as the input into our classifier for the upcoming phase.

B. Bidirectional networks

Having obtained feature maps of all sEMG signals, BLSTM is a proper classifier to realize classification. LSTM units were firstly came up with by Hochreiter and Schmidhuber to solve gradient vanishing problem in Recurrent Neural Networks (RNN) [21], and they were particularly appropriate for longer sequences processing. Based on the RNN model, LSTM units introduce a self-adapting gating mechanism, which can



(a) Classifier structure 1

(b) Classifier structure 2

Fig. 4. The overview of the MSR classifier. (a) and (b) denote two classifier structures, with a difference in attention layer's position. Their classification results are compared in Section IV. Every step's output size is shown in the figure.

memorize previous states selectively in a better way.

$$\begin{aligned}
i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + b_i\right)\\
z_t &= \operatorname{softsign}\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right)\\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right)\\
o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + b_o\right)\\
c_t &= f_t \odot c_{t-1} + i_t \odot z_t\\
h_t &= o_t \odot \operatorname{softsign}\left(c_t\right)
\end{aligned} \tag{3}$$

In the equations above, i_t, f_t, o_t serve as input gate, forget gate, output gate, respectively. z_t acts as a concatenated input state. The W represents the weight matrixes, and parameters b are the bias vectors. σ means the logistic sigmoid function and $\odot$ denotes Hadamard product. Similar to RNN, the output of the LSTM unit is also obtained by transforming hidden state h_t.

In this study, we use the bidirectional structure in LSTM units to conduct MSR. Bidirectional LSTM's capture both the previous timesteps (past features) and the future time steps (future features) via forward and backward states respectively [22]. By concatenating forward and backward hidden states, we can form a new hidden vector as:

$$h_t = \left[\overrightarrow{h}_t, \overleftarrow{h}_t\right] \tag{4}$$

Then the hidden vector feeds to the next layer or being converted to the output. The classifier structure is illustrated in "Fig. 4(a)". We stack four layers of BLSTM networks to reinforce the MSR effect.

C. Self-attention mechanism

In order to focus more on significant regions, we introduce an attention layer in multi-layer BLSTM networks. With an essence of weighted feature selection, attention mechanism has been proven successful in encoder-decoder models, while self-attention mechanism extends its boundary to classification tasks [23]. In this work, we design a simple self-attention layer after a BLSTM block. The transform process is shown as follows:

$$\begin{aligned}
q_{ti} &= \tanh\left(W_a h_i\right)\\
\beta_{ti} &= \frac{\exp\left(q_{ti}^T q_a\right)}{\sum_i \exp\left(q_{ti}^T q_a\right)}\\
h_t^o &= \sum_i \beta_{ti} h_i
\end{aligned} \tag{5}$$

First, we feed the input h_i that the BLSTM layer produced through a one-layer MLP to get q_{ti} as a hidden representation of h_i [24]. The W_a and q_a are trainable weight matrices. Next, we employ softmax function to find the normalized attention weight β_{ti}. Then we use attention weights to weight the inputs linearly and get the output vector h_t^o. Finally, we can acquire the final output result $H^o = [h_1^o, h_2^o, \ldots, h_t^o, \ldots, h_n^o]$ and feed it to the layer below.

D. Multi-category loss function

As for the loss function of the classifier, we choose the Focal Loss (FL), which is greatly suitable for our dataset. Proposed by FAIR group in 2017, FL performs well when classification task facing the challenge of the imbalance number between hard-classified samples and easily-classified samples [25]. The original FL was used for binary classification, but we can write the multi-category FL as:

$$mcFL = -\sum_{i=0}^{C-1} y_i \alpha_i \left(1 - p_i\right)^{\gamma} \log\left(p_i\right) \tag{6}$$

This loss function can make the model focus more on difficult-to-classify samples during training by dynamically changing the loss weights. Every p_i in probability distribution $p = [p_0, p_1, \ldots, p_i, \ldots, p_{C-1}]$ denotes the probability of the sample belongs to the class i. The $y = [y_0, y_1, \ldots, y_{C-1}]$ is one-hot representation of labels. The α serves as the weighting factor, controlling weights in different categories, while γ is the tunable focusing parameter and C is the total class number.

IV. Experimental Results and Discussion

After signal preprocessing, STFT, and CNN feature extraction, we acquire a $(6, 1000)$ size feature map in each sample. Then we randomly divide the training set and the test set according to the ratio of 7:3. The training set contains 4200 samples while the test set involves 1800 samples. Next, we use multi-layer attention BLSTM and other classical classifiers to carry out MSR. Results confirm that we can utilize deep learning frameworks effectively to implement sEMG recognition without manual designing features. After building the recognition model by the training set, we realize the real-time control system, including real-time sEMG acquisition, real-time transmission, real-time processing and real-time recognition. The response time is less than 200ms, which confirms our system is applicable in real world situations.

A. Comparison of classification results

To reduce the random error, every classifier was tested five times and the average was taken as the final result. The MSR test set results are illustrated in “TABLE II”. The attention BLSTM method achieves highest test set accuracy in 0.9711, while BLSTM follows closely. It is clear that bidirectional networks boost the performance greatly, and self-attention mechanism also contributes a slight improvement. One of the prediction distributions based on the attention BLSTM networks are shown in the “Fig. 5”. F1-scores of the four words are very close, from 0.96 to 0.98. A relatively confusing pair of words is “Er” and “Kuai”.

TABLE II
Recognition Accuracy of Different Classifiers

Classifiers	**Accuracy** (Mean ± SD)	**F1-score** (Mean ± SD)
LSTM	0.8989 ± 0.0206	0.8984 ± 0.0209
BLSTM	0.9634 ± 0.0028	0.9630 ± 0.0029
Random Forest	0.9097 ± 0.0035	0.9106 ± 0.0033
CNN	0.9022 ± 0.0042	0.9024 ± 0.0043
Att-BLSTM	**0.9711 ± 0.0030**	**0.9706 ± 0.0028**

We also compare the effect concerning different positions of self-attention layer in multi-layer BLSTM model. The structure in “Fig. 4(a)” achieved 0.9711 accuracy and the structure in “Fig. 4(b)” is only 0.0062 lower. The change is not significant.

B. Limitations

As a developing technology, the application of sEMG-based MSR is surely subjected to some limitations right now. Compared with acoustic signal based speech recognition, the robustness of MSR is not high enough, and there are not many words that can be recognized. Therefore, until now, the most suitable application scenario for MSR is sending listed commands to remote control systems. In addition, compared with Random Forest, deep learning frameworks spend more time training models, and their real-time response time is relatively longer when working on a low-performance processor.

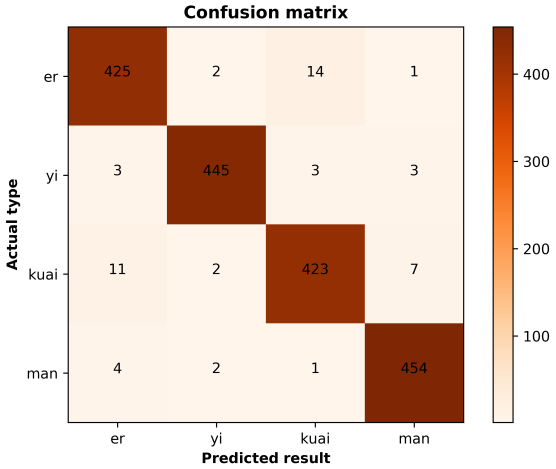


Fig. 5. One of the prediction distribution of attention BLSTM results. The number indicates the number of test set samples belonging to each situation. The F1-scores of each class range from 0.96 to 0.98.

V. Conclusion

This paper presents an automatic MSR system, which includes portable devices, STFT, CNN and multi-layer attention BLSTM model. STFT, together with CNN, serve as an automated feature extractor. Bidirectional networks plus attention mechanism improve recognition results remarkably. Moreover, by introducing deep learning methods, MSR no longer needs handcrafted signal features like ZC and RMS, reducing the effort of trying kinds of features. Thus, the whole recognition process is highly automatic.

Also, this study combines precision and convenience, representing a vital step towards MSR-based human-machine interaction. Our future research goal is to use the system for robotic arm control after expanding the corpus.

References

[1] M. Kim, W. K. Chung, K. Kim, “Preliminary Study of Virtual sEMG Signal-Assisted Classification,” IEEE 16th International Conference on Rehabilitation Robotics (ICORR), Toronto, pp. 105-110, 2019.

[2] F. Grandori, et al., “Multiparametric analysis of speech production mechanisms,” IEEE Eng. Med. Biol., vol. 4–5, no. 2, pp. 203–209, 1994.

[3] T. Schultz, et al., “Biosignal-based spoken communication: A survey,” IEEE/ACM Trans. Audio Speech Lang. Process., vol. 25, no. 12, Dec. 2017.

[4] S. Morse, Y. N. Gopalan, M. Wright, “Speech recognition using myoelectric signals with neural networks,” Proc. 13th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vol. 13, No. 4, pp.1877-1878, 1991.

[5] S. Kumar, D. K. Kumar, M. Alemu, and M. Burry, “EMG based voice recognition,” in Proc. IEEE ISSNIP, pp. 596–597, 2004.

[6] A. D. C. Chan, et al., “Multiexpert automatic speech recognition using acoustic and myoelectric signals,” IEEE Trans. Biomed. Eng., vol. 53, no. 4, pp. 676–685, Apr. 2006.

[7] A. D. C. Chan, et al., “Hidden Markov model classification of myoelectics signals in speech,” IEEE Eng. Med. Biol., vol. 9–10, no. 5, pp. 143–146, Sep./Oct. 2002.

[8] Kubo T, Yoshida M, Hattori T, Ikeda K. “Towards excluding redundancy in electrode grid for automatic speech recognition based on surface EMG,” Neurocomputing, 134:15–9, 2014.

[9] Lee KS. “EMG-based speech recognition using hidden Markov models with global control variables,” Trans. Biomed. Eng. 55(3):930–40, 2008.

[10] Q. Ai, et al., “Convolutional Neural Network applied in mime speech recognition using sEMG data,” Chinese Automation Congress (CAC), 2019.

[11] X. Cao, et al., “Gesture Recognition Based on ConvLSTM-Attention Implementation of Small Data sEMG Signals,” Adjunct Proceedings of UbiComp/ISWC, pp. 21-24, 2019.

[12] Jong, et al., “A speech recognition system based on electromyography for the rehabilitation of dysarthric patients: a Thai syllable study,” Biocybernetics and Biomedical Engineering, 39(1), pp. 234–245, 2019.

[13] G. S. Meltzner, et al., “Silent Speech Recognition as an Alternative Communication Device for Persons With Laryngectomy,” IEEE Trans. Audio Speech Lang. Process., vol. 25, 2017.

[14] W. Wei, W. Geng, et al. “A multi-stream convolutional neural network for sEMG-based gesture recognition in muscle-computer interface,” Pattern Recognition Letters, 2017, vol. 119, pp. 131-138.

[15] M. Zhang, “Inductive conformal prediction for silent speech recognition,” Journal of Neural Engineering, 2020.

[16] M. Wand, M. Janke, and T. Schultz, “Tackling speaking mode varieties in EMG-based speech recognition,” IEEE Transactions on Biomedical Engineering, vol. 61, no. 10, pp. 2515–2526, 2014.

[17] D. Phonetics, “Dissection of the speech production mechanism,” tech. rep., Working Papers in Phonetics, UCLA (102), 1–89, 2002.

[18] Richard L. Drake, Gray’s Anatomy. Beijing: Peking University Medical Press, 2010.

[19] D. Guo, et al., “A hybrid feature model and deep learning based fault diagnosis for unmanned aerial vehicle sensors,” Neurocomputing, vol. 319, pp. 155-163, 2018.

[20] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800-1807, 2017.

[21] S. Hochreiter, J. Schmidhuber, “Long short-term memory,” Neural Computation, 9(8):1735–1780, 1997.

[22] Guineng Zheng, Subhabrata Mukherjee et al., “OpenTag: Open Attribute Value Extraction from Product Profiles,” Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2018.

[23] Peng Zhou, Wei Shi, et al., “Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification,” Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 207–212, 2016.

[24] Z. Yang et al., Hierarchical Attention Networks for Document Classification, Proceedings of NAACL-HLT, pp. 1480–1489, 2016.

[25] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” IEEE International Conference on Computer Vision (ICCV), pp. 2999–3007, 2017.