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

Parallel-Inception CNN Approach for Facial sEMG based Silent Speech Recognition

Jinghan Wu, Tao Zhao, Yakun Zhang*, Liang Xie, Ye Yan, and Erwei Yin

Abstract— **With the purpose of providing an external human-machine interaction platform for the elderly in need, a novel facial surface electromyography based silent speech recognition system was developed. In this study, we propose a deep learning architecture named Parallel-Inception Convolutional Neural Network (PICNN), and employ up-to-date feature extraction method log Mel frequency spectral coefficients (MFSC). To better meet the requirements of our target users, a 100-class dataset containing daily life-related demands was designed and generated for the comparative experiments. According to experimental results, the highest recognition accuracy of 88.44% was achieved by proposed recognition framework based on MFSC and PICNN, exceeding the performance of state-of-the-art deep learning algorithms such as CNN, VGGNet and Inception CNN (3.22%, 4.09% and 1.19%, respectively). These findings suggest that the newly developed silent speech approach holds promise to provide a more reliable communication channel, and the application scenery of speech recognition technology has been expanded at the same time.**

I. Introduction

The aging of population is becoming a social issue in some countries, including China. With the development of technologies, we are now seeking the implementation of human machine interface in living assistance for the elderly. While automatic speech recognition (ASR) has been a mature technique for human machine interaction, it has some unavoidable limitations regarding to its poor performance in highly noisy environment. It is also inconvenient or even inaccessible for the elderly with disorders like dysphonia [1]. To better serve the target users, and further broaden the application of speech recognition, we adopt the silent speech recognition (SSR), which is independent of acoustic signals, in this research. It provides a communication platform for the elderly in need with great potential in recognizing human speech using physiological modalities, among which surface Electromyography (sEMG) has been studied and applied in many researches [2].

Speech recognition using EMG signals has been proposed since the mid-1980s, when researches showed that surface Electromyography (sEMG) signals contained speech-related information and could be used as an alternative of acoustic signals [3, 4]. According to previous researches [5, 6], time domain features are commonly used in feature extraction of sEMG signals. While spectrum features like Mel-Frequency Cepstral Coefficients (MFCC) used in automatic speech recognition [7, 8, 9] and power spectrum-based features extracted from sEMG signals [10] are also adopted in the signal processing nowadays with the development of machine learning algorithms. They show strong representation ability in previous researches and have promising effectiveness in silent speech recognition systems. On the other hand, machine learning algorithms such as feedforward neural network [11], random forests [12] and support vector machine [13] have been implemented in speech recognition process to recognize human intentions. With the development of computational devices, some deep learning architectures like convolutional neural network (CNN) [8] and bidirectional long short-term memory (BLSTM) [14] are also applied in speech recognition tasks, showing outstanding and stable performance in previous studies.

As far as we know, most researches on sEMG based silent speech recognition have been carried out for English with a few for Chinese language. Considering the datasets used in former studies, the complexity of sEMG signals collected by our own, and the differences among languages, our research focused on 100-class of Chinese phrases, and proposed a novel sEMG based silent speech recognition system using up-to-date feature extraction method and deep learning algorithm. Series of experiments on convolutional network based deep learning architectures and different features extraction methods were conducted. By being tested on designed dataset, our research provided a communication platform for the elderly, especially those with speech disorders. The main contributions of our research include：

i. Proposed a novel deep learning architecture named Parallel-Inception CNN for large dataset classification and recognition, and achieved satisfying performance.

ii. Extracted log Mel frequency spectral coefficients (MFSCs) from sEMG signals and carried out comparisons with several commonly used features.

iii. Designed the largest Chinese dataset used in sEMG based silent speech recognition studies, containing 100 classes of daily-life related demands for target users, and collected sEMG data from fourteen subjects.

iv. Extended the application scenarios of sEMG based silent speech recognition.

*Research supported by Natural Science Foundation of China (grant 62076250).

Jinghan Wu and Tao Zhao are with Tianjin University and Tianjin Artificial Intelligence Innovation Center (TAIIC), Tianjin, China.

Yakun Zhang, Liang Xie and Erwei Yin are with Defense Innovation Institute, Academy of Military Sciences (AMS), Beijing and Tianjin Artificial Intelligence Innovation Center (TAIIC), Tianjin, China (Yakun Zhang is the corresponding author to provide e-mail: ykzhang1222@126.com).

Ye Yan is with Defense Innovation Institute, Academy of Military Sciences (AMS), Beijing, China.

II. Materials and Methods

A. Dataset Design and Acquisition

We designed a dataset for our target users based on four kinds of demands in daily life. There were three to five Chinese characters in each demand (eighteen pieces containing three characters, forty-eight pieces containing four characters and thirty-four pieces for five characters, with an empty demand added in the end of the corpus when collecting sMEG signal data), covering a total of 190 different Chinese characters. Detailed contents could be divided into four aspects, from physiological demands (such as *I'm hungry*) to medical care requirements like *I need to take medicine* and *my leg aches*. We took some daily entertainments (e.g., *I want to listen to music*) into consideration, as well as social demands like *send text message for me*. As far as we knew, our dataset was the largest Chinese dataset for sEMG based silent speech recognition studies, and it was expected to cover as much potential needs for the elderly as possible.

Based on the contents of designed dataset, we carried out the data acquisition experiments. The sEMG data was captured from six positions of the facial and neck speech production related muscles, i.e., mentalis, risorius, levator labii superioris, anterior belly of the digastric, myhyoid and platysma. The device we used for sEMG signal amplification and acquisition was NSW308M bipolar EMG system with disposable Ag/AgCl surface electrodes collecting physiological signals.

Figure 1. Surface electrode used to collect sEMG signal data and display of electrode positions

A screen randomly displayed a phrase in the corpus and sustained for 2 seconds when collecting data. In particular, the subjects uttered Chinese phases in corpus, and the data was collected from fourteen (seven male and seven female subjects, aged from 23 to 28, with a mean age of 24.72) normal subjects manually, with Mandarin as their mother language. Each subject was asked to utter the phrase in subvocal mode during the period that it was indicated on screen, and the experiment system recorded and saved the real-time sEMG signals from the movement of those muscles simultaneously. The whole experimental procedure was approved by the Institutional Review Board. The informed consent form was given to each subject and signed before experimental procedure. All the phrases in the corpus were spoken once in a session. There were 140 sessions in total, including 14,000 pieces of sEMG data, as all the subjects were required to repeated 10 sessions.

B. Data Pre-processing

After the online signal collection, the raw sEMG data was further recorded at the sampling rate of 1000 Hz. We took some simple filters to improve the signal to noise ratio (SNR) before sEMG data could be used for feature extraction and recognition. A Butterworth notch filter of 50 Hz was first adopted for 50 Hz power frequency noise removal. As the main frequencies of effective sEMG signals were distributed at the range of 5-500 Hz, especially in the range of 10-400 Hz [15], we used a Butterworth bandpass filter with 10-400 Hz to obtained the most effective part of the sEMG signals.

C. Silent Speech Recognition System

Fig 2 shows a block diagram of proposed sEMG-based Chinese silent speech recognition system. In this study, the recognition system was consisted of five components, including sEMG data acquisition, data pre-processing, feature extraction, model training with labels and recognition in testing. We studied and tried different feature extraction methods, as well as recognition algorithms to improve the overall performance of our speech recognition system, carrying out comparative experiments and selected the best solution for our original purpose of providing a human-machine interface for the elderly in need. Detailed information for these two parts is demonstrated below.

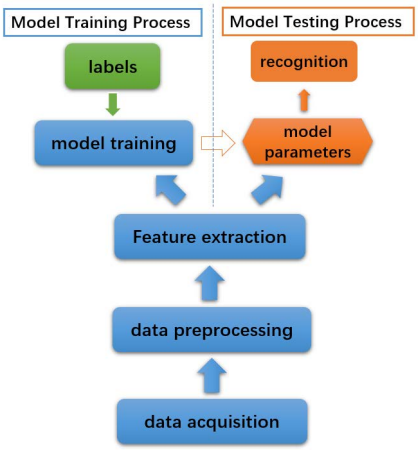

Figure 2. Pipeline of silent speech recognition system

D. Feature Extraction Methods

Surface electromyography signal is non-stationary and stochastic, varying greatly over time. Therefore, it is an essential step to extract short-term features from sEMG signals before modelling and recognition. We dropped the beginning 250 ms of the filtered sEMG signals as a reaction period of each subject. A sliding time window with fixed length of 200 ms was used to extract four of time and frequency domain features, i.e., Mean Absolute Value (MAV), Variance (VAR), Mean Frequency (MNF) and Wavelet Transform (WT). It moved 50 ms forward each time, thus we obtained a dimension of 32 for each feature and these four features were normalized and combined to form a new feature vector named TFD4 with dimension of 128 for each channel of sEMG signals.

Apart from TFD4, Mel frequency cepstral coefficient (MFCC) and log Mel frequency spectral coefficient (MFSC, also known as logarithmic filter-bank energies) would also be calculated as different measures of signal features. Mel filter-bank filtering was used in the extraction of these two features, considering the auditory characteristics of human.

TABLE I. TIME AND FREQUENCY DOMAIN FEATURES USED HERE

Feature	Basic formulation
Mean Absolute Value (MAV)	$MAV_k = \frac{1}{W}\sum_{i=1}^{W}\lvert x(i)\rvert$
Variance (VAR)	$VAR_k = \frac{1}{W}\sum_{i=1}^{W}\left(x(i) - \frac{1}{W}\sum_{i=1}^{W}x(i)\right)^2$
Mean Frequency (MNF)	$MNF_k = \frac{\sum_{i=1}^{W} f_i P_i}{\sum_{i=1}^{W} P_i}$
Wavelet Transform (WT)	$WT(a,\tau) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} f(t)\,\Psi\left(\frac{t-\tau}{a}\right)dt$
Parameter definition	*x(i)*: time series *k*: the *k*th time window W: length of sliding time window *fi*: value of the *i*th frequency P*i*: power spectrum density of the *i*th frequency *a* and τ: wavelet parameters

MFSC, on the other hand, omitted the DCT process in MFCC feature extraction and had relatively higher dimension. As far as we knew, it was the first time that MFSC features implemented in sEMG based silent speech recognition. Although it was originally proposed for acoustic speech recognition, we would like to confirm whether MFSC was appropriate to represent sEMG signal characteristics in this paper.

For MFCC features, we implemented 36 frames, 12 Mel filters, first order difference and second order difference of Mel filters to form a 36x36 input dimension. For MFSC, we implemented 36 Mel filters and 36 frames to obtain the input features of each channel.

E. PICNN Architecture

Convolutional neural network (CNN) has gained increasing popularity in image and signal processing tasks. The original architecture of CNN was firstly built by Yann LeCun in 1988, called LeNet [16], having 7 layers in total. With the development of deep learning algorithms, VGGNet [17] attracted much attention for its deep structure and outstanding performance in extracting features from the input data. The Inception (or GoogLeNet) [18] was developed by Google in 2015. It employed a new element called Inception module, from which we saw potential of implementation in sEMG based speech recognition. We came up with our own model based on Inception CNN, a novel architecture that designed and proposed in this paper, named Parallel-Inception CNN (PICNN).

In Parallel-Inception CNN, the input features were parallel processed based on their channels as can be seen in Fig 3. Filters with different sizes were employed and combined to form one inception module for each channel and perform convolution operation simultaneously. In each inception module, filters with size 1x1, 3x3 and 5x5 were used here and the number of filters with each size was 32. All the six modules performed convolution in parallel and then they were concatenated and put into the rest part of a common convolutional neural network. The idea of designing parallel convolution at the input of network was based on the fact that experimental data were collected by electrodes of six channels upon different facial muscles, and the distribution of the signal from each channel was mainly related to the muscle movements during speech. In parallel convolution, the parameters of convolution kernel used in each channel were different after training, which could obtain the data feature of each independent channel more effectively and improve the representation ability of the network.

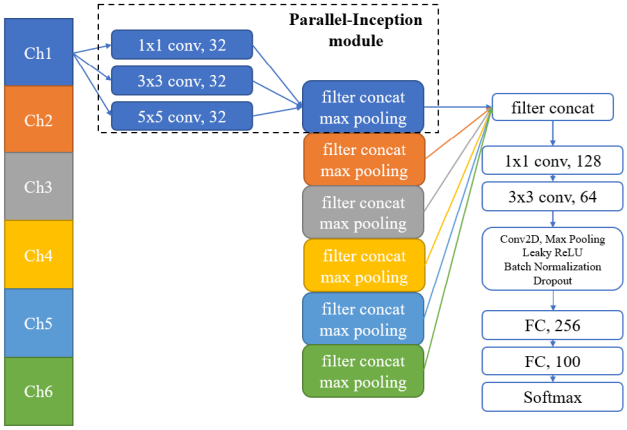

Figure 3. Architecture of proposed Parallel-Inception CNN

III. EXPERIMENTS AND DISCUSSIONS

A. Model Training Strategies

For more efficient model training, Mini-Batch Gradient Descent was adopted for the parameter update of neural networks during training process. Adam optimizer and learning rate decay were also implemented here for quicker convergence and more stable parameter refresh. Batch normalization had been added to each activation layer in proposed recognition model to handle the fluctuations of layer parameters during training and testing. Dropout was used to avoid overfitting, neurons with a percentage of 25% would be randomly dropped to improve the generalization of our model.

The whole dataset collected and pre-processed in above sections was shuffled randomly. Then 5-fold cross-validation was applied. When the recognition rate kept steady after 500 epochs of training, we stopped the process.

B. Experiment Results and Discussions

The average recognition accuracies using different features were recorded in Table II. Recognition system with MFCC or MFSC features achieved 5.58% and 19.2% higher classification accuracy, compared to that using TFD4 respectively, which indicated that MFCC and MFSC features kept more original characteristics and trends in raw sEMG data. On the other hand, MFSC features performed better than MFCC features and reached the best recognition accuracy of 88.44%, which proved the effectiveness of MFSCs in sEMG based silent speech recognition.

TABLE II. RECOGNITION ACCURACY USING DIFFERENT FEATURES

Features	Training accuracy	Testing accuracy
	Using Parallel-Inception CNN model	
TFD4	73.03	69.24
MFCC	89.76	74.82
MFSC	96.79	**88.44**

Apart from PICNN, three CNN architectures were tested, i.e., original architecture (denoted as CNN), Inception CNN and VGG16, and the results were recorded in Table III. The implementation of parallel inception module was proved to be more efficient in recognition and classification, with a rise of 4.09% in recognition accuracy than using VGG16. According to the accuracy summary in two experiments, the proposed Parallel-Inception CNN was more suitable for speech recognition on the large dataset we collected and kept much space for further study, according to the experiment results listed here.

TABLE III. RECOGNITION ACCURACY USING DIFFERENT DEEP LEARNING ALGORITHMS

Classifier	Training accuracy	Testing accuracy
	Using MFSC features	
CNN	97.94	85.22
VGG16	98.50	84.35
Inception	98.47	87.25
Parallel-Inception CNN	96.79	**88.44**

During dataset generation, we also noticed the physiological differences among subjects. We compared all the 14 subjects from aspect of gender in Table IV. Interestingly, a better recognition performance was achieved by using data collected from male subjects. As the subjects were of similar ages and health conditions, further analysis would be conducted from individual aspects to see how they affected the recognition result.

TABLE IV. RECOGNITION DIFFERENCES REGARDING SUBJECT GENDER

Gender (Number of subjects)	Training accuracy	Testing accuracy
	Using MFSCs and Parallel-Inception CNN	
Male (7)	98.37	90.95
Female (7)	97.85	86.93

IV. CONCLUSIONS AND FUTURE WORK

Our newly designed Parallel-Inception CNN (PICNN) model achieved an average recognition accuracy of 88.44% for a 100-class dataset collected by our own in this paper. The performance of proposed PICNN surpassed state-of-the-art convolutional architectures on our dataset. By being tested with different feature extraction methods, the best recognition result of our silent speech recognition system was achieved.

We are going to enlarge our dataset by inviting more aged subjects and patients to participate in the data collection. The dataset will be processed for more researches and open-access in the future. Also, cross-subject experiments will be carried out to further validate the robustness of proposed method and analyze the differences brought by subjects.

REFERENCES

[1] Green, Phil, et al. "Automatic speech recognition with sparse training data for dysarthric speakers." Eighth European Conference on Speech Communication and Technology. 2003.

[2] Schultz, Tanja, et al. "Biosignal-Based Spoken Communication." IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 12, 2017, pp. 2257-2271.

[3] Sugie, Noboru, and Tsunoda, Koichi. "A Speech Prosthesis Employing a Speech Synthesizer-Vowel Discrimination from Perioral Muscle Activities and Vowel Production." IEEE Transactions on Biomedical Engineering, no. 7, 1985, pp. 485–490.

[4] Morse, Michael S., and Edward M. O'Brien. "Research summary of a scheme to ascertain the availability of speech information in the myoelectric signals of neck and head muscles using surface electrodes." Computers in biology and medicine, vol. 16, no. 6, 1986, pp. 399-410.

[5] Srisuwan, Niyawadee, et al. "Comparison of Feature Evaluation Criteria for Speech Recognition Based on Electromyography." Medical & Biological Engineering & Computing, vol. 56, no. 6, 2018, pp. 1041-1051.

[6] Mendes Junior, José Jair Alves, et al. "Analysis of Influence of Segmentation, Features, and Classification in SEMG Processing: A Case Study of Recognition of Brazilian Sign Language Alphabet." Sensors (Basel, Switzerland), vol. 20, no. 16, 2020, p. 4359.

[7] Zhang, Ming, et al. "Inductive conformal prediction for silent speech recognition." Journal of Neural Engineering, 2020.

[8] Kapur, Arnav, et al. "AlterEgo." 23rd International Conference on Intelligent User Interfaces, 2018, pp. 43-53.

[9] Meltzner, Geoffrey, et al. "Silent Speech Recognition as an Alternative Communication Device for Persons With Laryngectomy." IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 12, 2017, pp. 2386-2398.

[10] Al-Timemy, Ali H, et al. "Improving the Performance Against Force Variation of EMG Controlled Multifunctional Upper-Limb Prostheses for Transradial Amputees." IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 24, no. 6, 2016, pp. 650-661.

[11] Jong, N. S., M. Kiatweerasakul, and P. Phukpattaranont. "Channel Reduction in Speech Recognition System based on Surface Electromyography." 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2018.

[12] Zhang, M., et al. "Feature selection of mime speech recognition using surface electromyography data." Chinese Automation Congress (CAC), 2019.

[13] Rameau, Anaïs. "Pilot Study for a Novel and Personalized Voice Restoration Device for Patients with Laryngectomy." Head & Neck, vol. 42, no. 5, 2019, pp. 839-845.

[14] Ye, H., et al. "Attention Bidirectional LSTM Networks Based Mime Speech Recognition Using sEMG Data." IEEE SMC 2020, 2020.

[15] Politti, Fabiano, et al. "Characteristics of EMG frequency bands in temporomandibullar disorders patients." Journal of Electromyography and Kinesiology, 2016, pp. 119-125.

[16] Lecun, Y., et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE, vol. 86, no. 11, 1998, pp. 2278-2324.

[17] Simonyan, Karen, and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[18] Szegedy, Christian, et al. "Going Deeper with Convolutions." IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1-9.