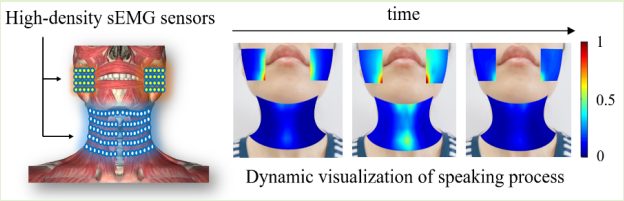
Automatic Speech Recognition in Different Languages Using High-Density Surface Electromyography Sensors

Mingxing Zhu, *Graduate Student Member, IEEE*, Zhen Huang, Xiaochen Wang, *Graduate Student Member, IEEE*, Xin Wang, *Associate Member, IEEE*, Cheng Wang, *Graduate Student Member, IEEE*, Haoshi Zhang, Guoru Zhao, *Member, IEEE*, Shixiong Chen, *Member, IEEE*, and Guanglin Li, *Senior Member, IEEE*

***Abstract*—Automatic speech recognition (ASR) based on surface electromyography (sEMG) sensors is an important technology converting electrical signals into computer-readable textual messages, which can overcome the limitation of acoustic sensors that are easily contaminated by environmental noises. However, current placements of sEMG sensors mainly depend on the experimenter's experience, which could miss important information about the major muscular activities and lead to the decline of classification performance. In this study, 120 closely-spaced sEMG sensors were utilized to collect high-density sEMG signals for recognizing ten digits in English and Chinese. The linear discriminant analysis classifier was used to classify the speaking tasks, and the sequential forward selection algorithm was utilized for analyzing the optimal position of the sensors. The results showed that the HD sEMG energy maps could help visualize the dynamic muscle activities during the speaking process, and significantly different muscular contraction patterns were observed for different speaking tasks. The classification accuracies when using the facial sensors were significantly lower than those on the neck, although with the same number of sensors. Moreover, the classification rates could be higher than 90% with only 15 optimally selected sensors that were mainly distributed on the neck instead of the face. This study suggests that the neck muscles could be the main contributor, and more sEMG sensors should be placed on the neck to improve the ASR performance. The findings of this study could provide valuable clues for the development of a practical sEMG-based speech recognition system, especially for patients with speaking disorders.**



***Index Terms*— Automatic speech recognition, high-density surface electromyography, sensors placement, sequential forward selection algorithm.**

Manuscript received October 10, 2020; revised November 5, 2020; accepted November 6, 2020. Date of publication November 10, 2020; date of current version June 30, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 81927804, Grant 61771462, and Grant 61901464; in part by the Shenzhen Governmental Basic Research Grant under Grant JCYJ20180507182241622; in part by the Science and Technology Planning Project of Shenzhen under Grant GJHZ20190821160003734; in part by the Shenzhen Science and Technology Development Fund under Grant JCYJ20170818163505850; in part by the Science and Technology Program of Guangzhou under Grant 201803010093; in part by the Shenzhen Institute of Artificial Intelligence and Robotics for Society; and in part by the Science and Technology Planning Project of Guangdong Province under Grant 2019A050510033. The associate editor coordinating the review of this article and approving it for publication was Prof. Rosario Morello. *(Mingxing Zhu and Zhen Huang contributed equally to this work.) (Corresponding authors: Shixiong Chen; Guanglin Li.)*

Please see the Acknowledgment section of this paper for the author affiliations.

Digital Object Identifier 10.1109/JSEN.2020.3037061

I. Introduction

SPEAKING activity, as one of the necessary ingredients of human life, is an essential way for human social communication. Speaking is a complex process controlled by a large number of articulatory muscles associated with phonation. Speaking different words or languages requires different ways of pronunciation and therefore involves different muscular contraction patterns, which could be recorded by a non-invasive technique called surface electromyography (sEMG) via placing EMG sensors on the skin surface for measuring the corresponding electrical signals. Since the sEMG signals contain substantial dynamic information about the articulatory muscle activities, the sEMG sensors could be used in automatic speech recognition (ASR) systems that convert the electrical sEMG signals associated with human speaking into computer-readable textual messages [1].

Unlike conventional recognition methods using the human voice collecting from acoustic sensors, the sEMG-based ASR systems do not rely on any acoustic signals, that are not always available and easily contaminated by various environmental noises. Therefore, it can be used even if the subject does not produce any audible voices (silent speech), such as patients with speech disorders [2], [3]. Therefore, the sEMG-based ASR system has developed into a prevalent technique with a wide variety of applications for speaking recognition in both audible and silent modes [4]–[6].

Since the sEMG technique is non-invasive and easy to use, the sEMG-based ASR has been reported in numerous studies during the past decades. For example, fifteen English words were classified by using the sEMG signals recorded from two sensors over the neck muscles of the subject with a firefighter's self-contained breathing apparatus [7]. In another study, a three-channel EMG system was developed for patients with speech impairment, and three Arabic vowels were recognized by using the sEMG signals recorded from facial muscles [8]. Three channels of sEMG sensors were placed on the facial muscles, and eleven voiceless Bangla vowels were classified by using the artificial neural network [9]. A total of eight sEMG sensors (4 on the face and 4 on the neck) were used to record the sEMG signals when reading phrases constructed from a 2500-word vocabulary for silent speech recognition of patients at least 6 months after total laryngectomy [3]. Five sensors (two on the face and three on the neck) were used to acquire the sEMG signals, and fourteen sEMG features and four classifiers were examined to classify eleven Thai words [10]. Five channels of sEMG sensors located on the facial muscles were utilized to classify nine Thai syllables for the rehabilitation of dysarthric patients [11]. Ten sEMG sensors placed on the facial and neck muscles were used to recognize ten specific silent speech commands in Chinese [12]. Moreover, different sEMG-based speech recognition systems have also been developed for different languages, such as English [13]–[15], Chinese [16]–[18], Japanese [19], [20], Thai [21], [22], Korean [23], Aceh [24] and Malay [25].

However, the placements of the sensors in the above-mentioned studies were mostly decided based on the experience or the trial-and-error method of the experimenter without any quantitative analysis, leading to a possible declination of the performance of the ASR system due to improperly placed sensors. A possible solution might be that the experimenter could place the sensor according to the physical distribution of the articulatory muscles [26]. Nevertheless, the speaking process is complex neuromuscular activities involving a larger number of small facial and neck muscles, and therefore the speaking of different words might generate dramatically different patterns of muscular involvements [27]. Moreover, each language may have its unique activation pattern of the articulatory muscles because of its specific pronunciation style [28]. Studies showed that the role of the articulatory muscles could be significantly different for different languages, and the placement of sensors could considerably affect the performance of the ASR system accordingly [29], [30]. Therefore, the investigation of the contribution of different articulatory muscles is helpful for providing objective guidelines on optimal sensor placements in cases with an inadequate number of sensors, so that the accuracy of the sEMG-based ASR system could be considerably improved. However, up to the present time, there are few studies to investigate the contributions of the facial and neck muscles in speech recognition of different languages.

In addition, most of the previous studies used only a few sensors located on the facial and/or neck muscles to record sEMG signals as the input of the sEMG-based ASR system. However, the muscles responsible for speaking are characterized by a large number and small shapes, and these muscles spanned a relatively large area across the face and neck to achieve subtle movements. The usage of a few empirically placed sensors in sEMG measurements may not provide adequate information to investigate the contributions of the facial and neck muscles in speech recognition. It is still not clear the muscles of which region (the face or the neck) play a more important role in the ASR system, due to the lack of comprehensive analyses of full information from all the muscles. Thus, these challenges have motivated the emergence of the high-density sEMG (HD sEMG) technique using multi-channel sEMG sensors in the sEMG-based ASR field. The HD sEMG technique uses a large number of closely placed sensors to record electrical activities of a large area of muscles so that the comprehensive information of a group of target muscles could be fully revealed [31]. Over the past few decades, the HD sEMG signals had been adopted in many research studies to decode motion intents for human-machine interaction systems, to evaluate the swallowing functions in patients with dysphagia, to study the behavior of the paraspinal muscles in people with low back pain and to analyze the motor unit decomposition in a non-invasive way [32]–[34]. It is also clinically useful in the assessment of motor fiber conduction velocity [35] and fatigue evaluation of motor unit action potentials [36] due to its non-invasiveness and its capacity to record over very long periods. The introduction of HD sEMG technique into the sEMG-based ASR system could overcome the limitation of current methods with insufficient sensors so that complete information about articulatory muscles could be obtained to analyze the contributions of the facial and neck muscles in ASR, which helps to provide practical guidelines on how to place the sEMG sensors to improve the performance of the sEMG-based ASR system.

The purpose of this study is to investigate the contributions of different articulatory muscles for English and Chinese speech recognition using multi-channel sEMG sensors. A total of 120 surface sensors closely placed over the facial and neck muscles were utilized to simultaneously collect HD sEMG signals when the subjects were speaking ten English and Chinese digits, respectively. A set of topographic maps were constructed to visualize the dynamic energy distribution of the articulatory muscular activities during the speaking process. The classification accuracies were calculated and compared for different sensor groups of the face and neck regions. The distribution of the optimal sensors automatically selected by a sequential forward selection algorithm was also analyzed to investigate the roles of different articulatory muscles. This study could provide a useful guideline for appropriately placing sEMG sensors and pave the way for the development of a

TABLE I
SPEAKING TASKS OF TEN DIGITS IN ENGLISH AND CHINESE LANGUAGE

class	English	Pronunciation	class	Chinese	Pronunciation
1	zero	['zɪərəʊ]	11	零	[líng]
2	one	[wʌn]	12	一	[yī]
3	two	[tuː]	13	二	[èr]
4	three	[θriː]	14	三	[sān]
5	four	[fɔː]	15	四	[sì]
6	five	[faɪv]	16	五	[wǔ]
7	six	[sɪks]	17	六	[liù]
8	seven	['sevn]	18	七	[qī]
9	eight	[eɪt]	19	八	[bā]
10	nine	[naɪn]	20	九	[jiǔ]

clinically feasible system for sEMG-based speech recognition, especially for patients with speaking disorders.

II. MATERIALS AND METHODS

A. Subjects and Experimental Procedure

A total of eighteen healthy volunteers (eleven males and seven females) with normal speaking and hearing capabilities were recruited to participate in the experiment of this study. All of the volunteers were native Chinese speakers with no less than ten years of English learning. Before the speaking tasks, the subjects were introduced with the intentions and procedures of the experimental protocols in detail. The experiments were approved by the Institutional Review Board of Shenzhen Institutes of Advanced Technology (#IRB ID: SIAT-IRB-170815-H0178). Every subject willingly provided their written informed consent and permitted the scientific and educational use of their photos and data.

In the experiments, the subjects were required to speak different digits with an audible speech in both English and Chinese, and the corresponding HD sEMG signals were collected from the articulatory muscles on the face and neck regions by multi-channel sEMG sensors. Before each session, 40 seconds of electrical signals were recorded when each subject remained in a relaxed state without any speaking or movements to obtain the baseline for the sEMG signals. Then the subjects were asked to speak ten digits (0 to 9) in English and Chinese, respectively (Table I). For each trial, each digit was spoken within one second, followed by a three-second rest to avoid muscle fatigue. Each trial was repeated 28 times before continuing to the next digit. The experiments were carried out in an electromagnetic-shielded room to ensure high-quality HD sEMG recordings.

B. HD sEMG Acquisition

In this study, the HD sEMG signals were synchronously recorded by a total of 120 sEMG sensors closely placed on the face and neck regions. The REFA 128 system (TMSI, REFA, the Netherlands) was used for the data collection with a sampling of 2048 Hz for each channel. Before the data acquisition, the skin surface was cleaned carefully by using the alcohol pad for removing extra dust, dander, and skin oil that could affect the quality of the sEMG signals. The 120 sEMG sensors were arranged as a set of two-dimensional arrays to cover all the facial and neck muscles, and the distance between each adjacent sensor was kept at a small interval of 15 mm to obtain comprehensive electrophysiological information at a high spatial resolution. As shown in Fig. 1, eighty sensors were structured in a 5 × 16 grid evenly located on the neck muscles. Meanwhile, two sensor arrays in a 4 × 5 grid (40 channels in total) were symmetrically placed on the left and right sides of the facial muscles.

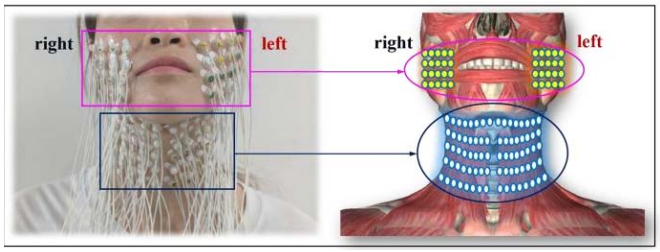


Fig. 1. Placement of the 120 sEMG sensors on the face and neck muscles for HD sEMG data acquisition.

In order to compare the contributions of different muscles in speech recognition, the sensor arrays were grouped in six ways (Fig. 2): (1) F-40: all the 40 sensors on the facial muscles (channel F1 to F40); (2) NO-40: the 40 sensors located at the odd columns of sensor arrays in the neck region (channel N1, N3, …, N79); (3) NC-40: the 40 neighboring channels located at the central area of the neck (channel N5 to N12, …, N69 to N76); (4) NE-40: the 40 sensors located at the even columns of the neck region (channel N2, N4, …, N80); (5) NA-80: all the 80 sensors on the neck muscles (channel N1 to N80); (6) FN-120: all of the 120 sensors on the facial and neck muscles.

C. HD sEMG Topographic Energy Maps

Firstly, the original sEMG data were filtered by a fourth-order Butterworth band-pass filter with cut-off frequencies from 30 to 500 Hz to attenuate low-frequency baseline wander and other high-frequency noises. Besides, a custom notch filter was utilized to reduce the power-line interferences at 50 Hz and its harmonic frequencies. Then, the HD sEMG signals of each channel were calculated by a set of analysis windows (length of 250ms) to generate the root mean square (RMS) of the HD sEMG recordings. Then, the RMS values were normalized (NRMS) across all channels of electrodes by using the maximum and minimum RMS of HD sEMG recordings. Afterward, a sequence of the topographic energy maps was constructed by the NRMS values for visualizing and evaluating the contraction patterns of the facial and neck muscles during the speaking tasks.

The RMS was calculated for each analysis window to obtain the average energy distribution of the muscular activities as

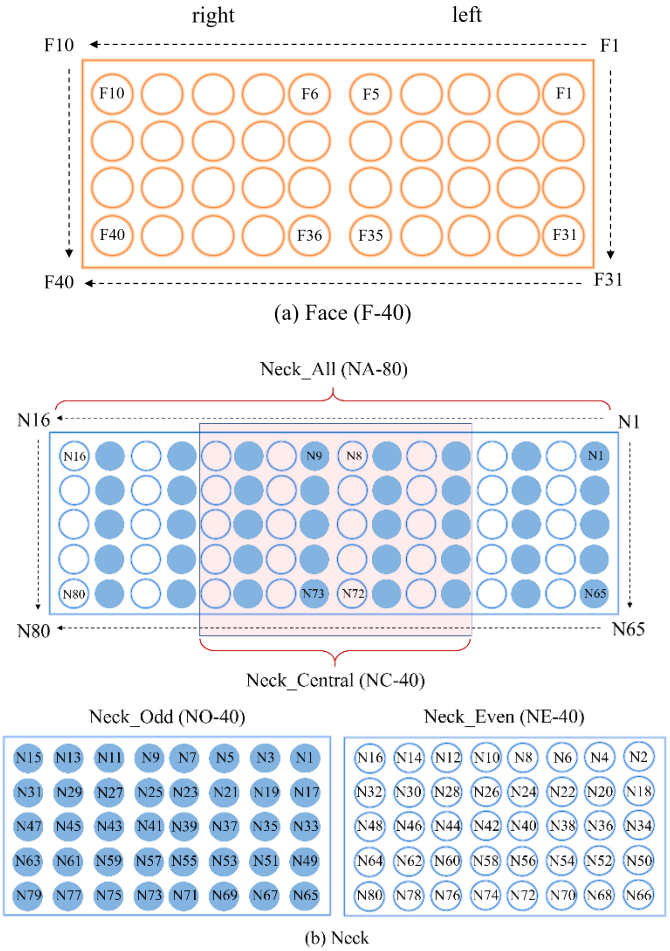


Fig. 2. Different ways to group the multi-channel sEMG sensors: (a) Face region: 40 sensors on the facial muscles (F-40). (b) Neck region: all the 80 sensors on neck muscles (NA-80), central columns (NC-40), odd columns (NO-40), and even columns (NE-40).

follows:

$$R\{v[m]\} = \sqrt{\frac{1}{m}\sum_{i=1}^{m} v^2[i]} \tag{1}$$

where $R\{v[m]\}$ is the RMS value of sEMG signals for each analysis window, $v[i]$ is the i^{th} sample in the analysis window, and m is the total number of windows.

The normalized RMS values were symbolized by NR as follows.

$$NR(i) = \frac{R(i,j) - \min(R)}{\max(R) - \min(R)} \tag{2}$$

where $NR(i)$ is the normalized RMS value of sEMG signals in channel i, $R(i, j)$ is the RMS value of channel i in analysis window j, $min(R)$ is the minimum RMS value of channel i, and $max(R)$ is the maximum RMS value of channel i.

D. Features Extracting and Word Classification

Then, the features of the HD sEMG signals were extracted for providing useful information embedded in the sEMG signals to recognize the intended speech tasks. The filtered HD sEMG signals containing all the 28 repetitions were manually sliced for each digit, with only the sEMG signals corresponding to the audible speaking process reserved to form the activity data. Afterward, the activity data containing the 28 repetitions of the same digit were partitioned into the sEMG series by a 400-point (almost 200 ms) sliding window with a 200-point increment for the computation of the sEMG features. Signal features that are in the time domain (TD), frequency domain (FD), and time-frequency domain (TFD) are used in sEMG-based pattern recognition. Among these different features in different domains, the TD features were used most frequently in sEMG classification due to their easy implementation, low computation complexity, and satisfactory performance [10], [37]–[41]. Moreover, Hudgins's feature set, including the Mean Absolute Value (MAV), Waveform Length (WL), Zero Crossing (ZC), and Slope Sign Change (SSC), could comprehensively reflect the temporal and spectral properties of sEMG signals [10] and therefore they were widely used by many other studies about prosthesis control and muscle-computer interface [42]–[47]. Thus, in this study, these four time-domain features, including the MAV, WL, ZC, and SSC, were extracted from the preprocessed sEMG signals for English and Chinese word classification, and the formula to compute these features were shown in Table II.

TABLE II
FOUR TIME DOMAIN FEATURES AND THEIR MATHEMATICAL DEFINITIONS

Feature (Abbr.)	Mathematical expression
Mean Absolute Value (MAV) [10, 48]	$MAV = \frac{1}{n}\sum_{k=1}^{n}\lvert x_k\rvert$
Waveform Length (WL) [10, 48]	$WL = \sum_{k=1}^{N-1}\lvert x_{k+1} - x_k\rvert$
Zero Crossing (ZC) [10, 48]	$ZC = \sum_{k=1}^{n-1}\left[\operatorname{sgn}(x_k * x_{k+1}) \cap (x_k - x_{k+1}) \geq Thr\right]$ $\operatorname{sgn}(x) = \begin{cases} 1, & if\ x \geq Thr \\ 0, & otherwise \end{cases}$
Slope Sign Change (SSC) [10, 48]	$SSC = \sum_{k=2}^{N-1}\left[f(x_k - x_{k-1}) * (x_k - x_{k+1})\right]$ $f(x) = \begin{cases} 1, & if\ x \geq Thr \\ 0, & otherwise \end{cases}$

x_k represents the EMG signal in a segment k and n denotes the length of the EMG signals.

Then 5-fold cross-validation arithmetic was employed to segment the matrix of extracted features and the corresponding targets into training and testing sets. These sets were subsequently fed into the linear discriminant analysis (LDA) classifier for recognizing the speech patterns inherent in the extracted sEMG features. Classification accuracy is one of the most popular metrics in various pattern recognition applications including speech recognition. In addition, classification accuracy is the simplest clustering quality measure to evaluate clustering results associated with the ground truth. It is essential for the accurate realization of a user's intent, and directly presents the recognition results of the speaking tasks. Thus, classification accuracy was considered as our core metric for evaluating the contributions of different articulatory muscles

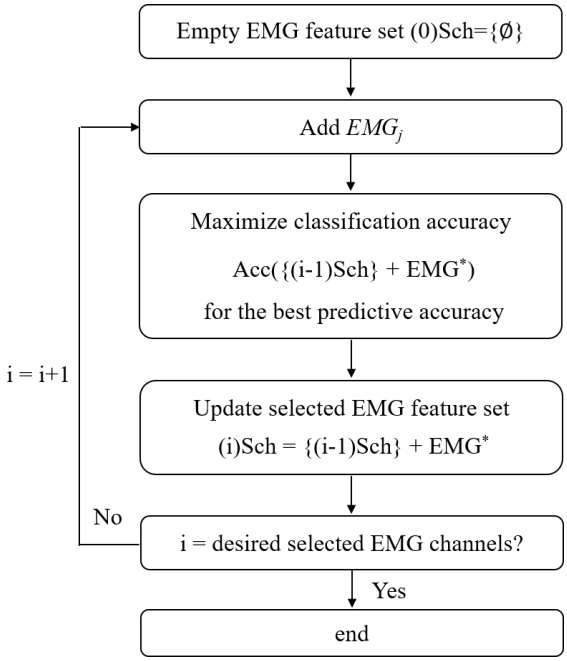


Fig. 3. Flowchart of sequential forward selection (SFS) algorithm to select the optimal sensors.

in speech recognition [49]:

$$Acc = \frac{Ncor}{Ntest} \times 100\% \tag{3}$$

where *Acc* is the classification accuracy, *Ncor* is the number of correctly classified samples, and *Ntest* is the total number of testing samples.

E. Sensor Optimization Analysis

In this study, the optimal sensor number was also calculated, and the distribution of the optimal sEMG sensors was analyzed to compare the contributions of different muscles for speech recognition. The sequential forward selection (SFS) algorithm, which automatically selects a subset of features that is most relevant to the problem, was employed to calculate the optimal sensor number for given classification accuracy. The SFS algorithm was easy to implement and shows great performance in various circumstances of data dimension reduction [50], [51]. The SFS algorithm started with a null feather set, and then the channel with the highest classification accuracy was selected among all the 120 channels. Subsequently, one more channel with the largest accuracy increment was added at each step of the algorithm until it reached a target desired classification accuracy, as shown in Fig. 3.

Given that the optimal channel sets *{(i-1)Sch}* containing a total of *(i-1)* channels had already been selected in the $(i\text{-}1)^{th}$ iteration for the SFS algorithm, each channel (EMG_j) from the rest sensors would be picked out and combined with the selected sets *{(i-1)Sch}* in the i^{th} iteration (4). This procedure was repeated until all the rest channels have been tested, and the optimal channel *EMG** with the highest classification accuracy would be selected for the i^{th} iteration. Accordingly, the sets *{(i-1)Sch + EMG*}* would be selected as the i^{th} optimal channel sets *{(i)Sch}* indicated by (5).

$$Acc\left(\{(i-1)\,Sch\} + EMG^*\right) = \max_{j\in\{1,2,\ldots,N-i\}} Acc\left(\{(i-1)\,Sch\} + EMG_j\right) \tag{4}$$

$$\{(i)Sch\} = \{(i-1)\,Sch\} + EMG^* \tag{5}$$

$$\{(0)Sch\} = \Phi \tag{6}$$

In this study, different numbers of optimal sEMG sensors, involving 5 channels (5-ch), 10 channels (10-ch), 15 channels (15-ch), 20 channels (20-ch), 25 channels (25-ch), and 30 channels (30-ch), were selected from the total 120 sEMG sensors by using the SFS algorithm, respectively. Then, the location of these optimally selected sensors was analyzed according to their distribution and the sensor number from different groups of muscles was counted separately.

The statistical analyses of one-way ANOVA were performed to analyze the effects of different sensor groups on the classification accuracies for Chinese and English speech recognition, respectively. Meanwhile, the distribution of the optimal sEMG sensors was also compared among different sensor groups to evaluate the contribution of different muscles for different speech recognition tasks. All the statistical results were obtained by comparing the p-value with a confidence level of 0.05. In this study, all the analyses of the offline HD sEMG data, such as digital filtering, feature extraction, SFS algorithms, and pattern recognition, were implemented in the Matlab software platform (MathWorks, Natick, MA, USA).

III. Results

A. HD sEMG Topographic Energy Maps for the Entire Speaking Process

In this study, the dynamic HD sEMG topographic energy maps, which could demonstrate the energy distribution of the articulatory muscular activities when the subject was speaking, were constructed from the sEMG signals and a typical example was shown in Fig. 4, where high energy intensity was represented by red color. The entire speaking process was segmented into six temporal frames (frame 1 to frame 6) for exhibiting the dynamic activities of the facial and neck muscles when the subject was speaking the English words “zero” and “one”, respectively.

Before the subject started to speak the word “zero”, the energy map kept at low intensity on both the face and neck regions in frame 1, as shown in Fig. 4(a). In frame 2, a high-energy area appeared at the bottom center of the neck, indicating the beginning of the word speaking. Then the energy concentration area started to move upward, and the maximum muscular activities were observed in the middle of the facial region in frame 3, with constantly diminishing EMG activities when moving away from the mouth. Afterward, the region with maximum muscular activities traveled downwards back to the lower edge location of the neck region, while the activities of the facial muscles decreased to a low intensity in frame 4. Thereafter, the intensity of the high-energy area on

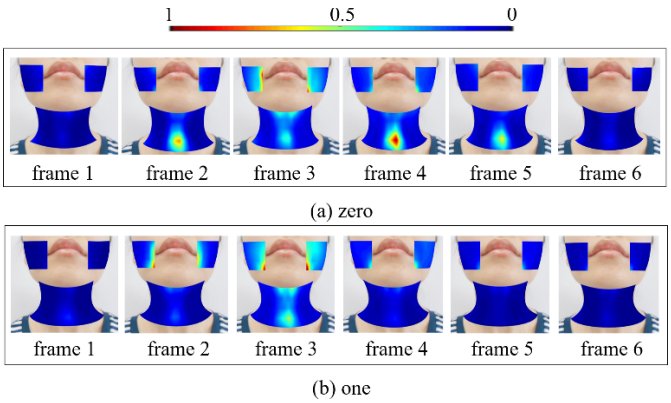


Fig. 4. Dynamic HD sEMG topographic energy maps during the entire speaking process when the subject was speaking "zero" (a) and "one" (b).

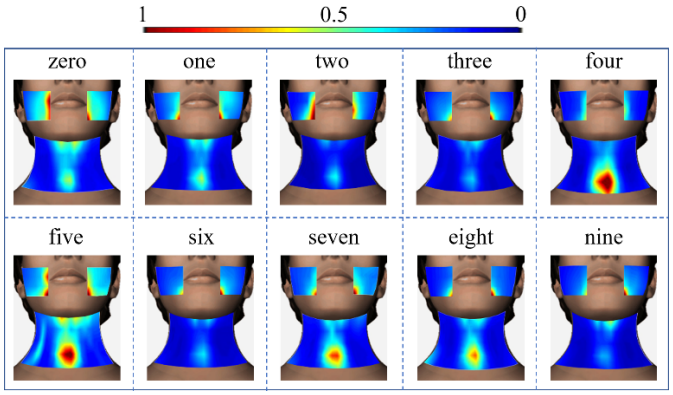


Fig. 5. The typical HD sEMG topographic maps when speaking ten different English words, including zero, one, two, three, four, five, six, seven, eight, and nine.

the neck gradually declined in frame 5, and finally disappeared in frame 6 when the speaking task completed.

On the contrary, the HD sEMG topographic energy map in Fig. 4(b) demonstrated a significantly different pattern when the subject spoke a different word of "One". Unlike Fig 3(a) in which the energy concentration area traveled forward and backward between the face and the neck, the EMG activities of the word "one" showed a briefer and simpler pattern. In Fig 3(b), noticeable muscular activities were first observed in frame 2 over the facial muscles around the mouth region. Then the intensity of the facial muscular activities considerably increased in frame 3, and the range of the active region spread downward to the center of the neck. After that, the intensity of the active areas significantly decreased in frame 4, with some residual energy distributed along the mouth region. From frame 5 to frame 6, no apparent muscular activities were observed on either the face or the neck region.

Additionally, when comparing with the two speaking tasks in Fig. 4(a) and 4(b), it was observed that the energy maps showed approximately symmetric left-and-right distributions for both the face and neck muscles during the whole speaking process.

B. Averaged HD sEMG Topographic Energy Maps for Different Words

For comparing the HD sEMG topographic energy maps among different speaking tasks, all the temporal frames (Fig. 4) during the speaking process were averaged for each digit word, and the averaged energy maps of 10 different words were shown in Fig. 5. It was observed that the EMG activities of the facial muscles were mainly located around the mouth regions while those of the neck muscles exhibited on the center of the neck across all the ten speaking tasks. Nevertheless, evident differences were also observed among different speaking tasks. While rather high intensities of muscular activities were observed on the neck region for the words "four", "five" and "seven", the significantly lower amplitude of neck energy distribution were seen for other words such as "two", "three", "six" and "nine". For the neck region, the area with the highest energy tended to locate at the lower portion for most of the word speaking tasks. Moreover, it was observed that the muscular activities showed coarse left/right symmetry for the neck regions. However, significant differences between the left and right could be observed in the facial region, especially for the words of "zero", "two" and "five". In other words, the facial areas with the highest energy were inclined to distribute around the mouth at the lower portions.

C. Comparison of Classification Accuracies Among Different Sensor Groups

To evaluate the performance of the speech recognition system among different sensor groups, the confusion matrices of classification accuracies were computed and compared for the F-40 and NO-40 sensor groups, as shown in Fig. 6. It was noted that the accuracy of the "rest" task attained 100% on F-40 and NO-40 groups for both English and Chinese recognition. In Fig. 6(a), for the F-40 group, while the accuracy reached 91.3 % for recognizing digit one, it dropped to around 67% in classifying digits eight and nine. Most of the accuracies were lower than 80% for the F-40 group. In contrast, most of the English words had a recognition accuracy above 80%, with the only exception of digit seven. In Fig. 6(b), using the F-40 sensor group for Chinese speech recognition showed slightly higher overall accuracy than English recognition tasks, with half of the tasks having accuracies higher than 80%. For the NO-40 group, only one task (digit 6) had a classification accuracy less than 80% and the highest accuracy could reach up to 94.5%.

For investigating the contributions of different muscle activities towards the sEMG based speech recognition, the 120 HD sEMG sensors were grouped in six different ways based on their locations (Fig. 2): F-40, NC-40, NO-40, NE-40, NA-80, and FN-120. A typical example of the classification accuracy (averaged across digits) as a function of the sensor group was shown for both languages in Fig. 7. It was observed that the F-40 and NC-40 groups had the lowest averaged classification accuracies (as low as 77.92%) for both languages. With the same number of sensors, the NO-40 and NE-40 groups showed significantly better performance with a classification accuracy as high as 91.58%, and there were no significant differences between the two groups. When all the 80 neck sensors were used for the recognition, the NA-80 group demonstrated the highest classification accuracy up to 95.09%. Moreover, the

	0	1	2	3	4	5	6	7	8	9	rest
0	78.4	1.0	0	3.5	5.1	2.3	0.9	0	6.2	1.7	1.0
1	0	91.3	3.9	0	0	0.5	0	1.3	0	3.0	0
2	0.7	0.9	70.5	0	2.6	0	2.6	0	18.4	4.4	0
3	4.1	0	0.6	70.6	0.4	0	7.1	2.1	10.4	4.7	0
4	0.8	1.0	1.6	1.2	83.2	1.9	2.7	0	4.4	2.9	0.4
5	4.1	0	1.6	0	6.0	77.4	0.5	0	7.6	2.8	0
6	0.4	0	0	2.9	0	0	70.1	2.8	16.3	6.0	1.5
7	0	0	0	1.5	0.6	0.7	7.1	83.3	2.1	4.7	0
8	1.0	0	0.5	7.0	4.5	0	8.5	0	67.7	10.8	0
9	0.4	0.7	2.6	5.1	1.6	0.8	2.0	0.6	19.0	67.4	0
rest	0	0	0	0	0	0	0	0	0	0	100

F-40

	0	1	2	3	4	5	6	7	8	9	rest
0	90.7	0.5	1.6	0	0	1.6	2.5	0	0	3.2	0
1	0	89.2	6.0	0	0.44	0	1.9	0	0	2.5	0
2	1.0	0	91.6	0	0	2.0	3.4	0	0	2.0	0
3	0.5	1.7	0	89.0	2.1	0	0.9	0.9	4.5	0.5	0
4	1.6	0.5	2.0	1.0	85.0	5.3	1.8	0.5	2.5	0	0
5	0.5	1.5	1.0	0.5	9.3	85.2	1.0	0	0.6	0.5	0
6	2.0	0.4	3.4	0	1.1	1.0	85.7	0	1.6	4.8	0
7	0	0.4	3.8	2.1	4.0	1.6	4.0	76.9	5.4	1.9	0
8	0	0.4	0	1.6	0.5	1.0	1.6	0.5	90.8	3.7	0
9	3.7	1.3	4.7	0	0	0.4	3.5	1.8	3.5	81.1	0
rest	0	0	0	0	0	0	0	0	0	0	100

NO-40

(a) English

	0	1	2	3	4	5	6	7	8	9	rest
0	86.2	1.8	2.9	0.4	0.4	0	1.8	5.2	0	1.4	0
1	4.8	75.0	0.4	0	1.1	0	1.1	17.7	0	0	0
2	8.2	0	90.2	0.4	0	0	0	0.5	0.4	0..4	0
3	10.0	1.0	1.1	84.9	1.5	0	0.3	1.1	0	0	0
4	6.1	0.3	0	0.3	87.4	0	1.4	3.7	0	0.8	0
5	9.1	0	2.3	0.3	0	69.5	8.3	0.9	0	9.6	0
6	6.9	0.4	0.4	0	0	0	78.6	4.6	0	9.2	0
7	4.6	8.9	0	0.4	0	0	1.2	84.2	0	0.8	0
8	1.7	0	2.2	09.7	0	0	1.9	1.1	78.0	0.7	4.7
9	2.6	1.4	0.7	0	1.5	0	9.3	5.5	0	79.2	0
rest	0	0	0	0	0	0	0	0	0	0	100

F-40

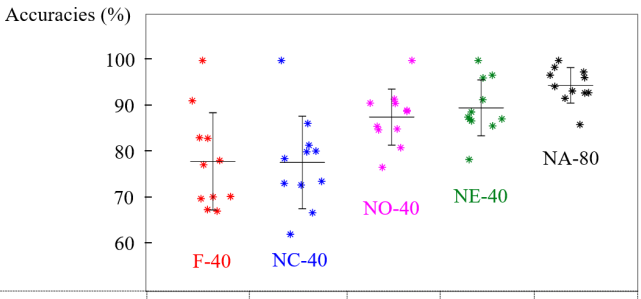
	0	1	2	3	4	5	6	7	8	9	rest
0	88.8	2.6	0.4	0	0	0	7.2	0.4	0.6	0	0
1	9.1	87.3	0.4	0	1.5	0	0.7	1.1	0	0	0
2	4.2	0	94.5	0	0	0	0.5	0	0.9	0	0
3	4.1	0.3	0	88.3	0.8	0	0.3	3.4	1.6	1.1	0
4	1.5	0.4	0	0	94.2	0	1.8	0.7	0	.14	0
5	1.7	0	0	1.5	1.8	84.2	4.3	0	1.1	5.5	0
6	8.2	0	0	0.4	8.7	0	79.2	1.2	0.4	2.1	0
7	0.4	1.6	0	0.7	1.1	0.3	1.5	93.8	0	0.7	0
8	2.5	0	0	5.8	0	0	0.7	0.7	87.0	3.4	0
9	0.4	0	0	0.8	5.2	0	2.7	2.1	0.8	88.2	0
rest	0	0	0	0	0	0	0	0	0	0	100

NO-40

(b) Chinese

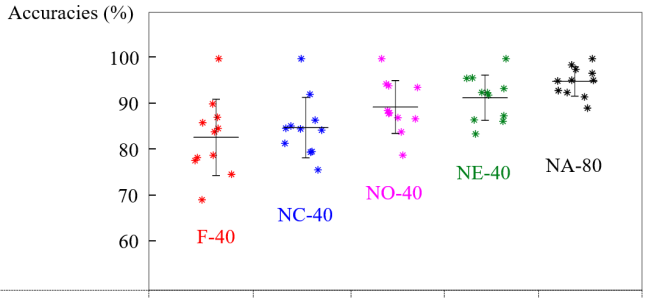
Fig. 6. The confusion matrixes of the classification accuracies using different sensor groups (F-40 and NO-40) for English (a) and Chinese (b) speech recognition.

Chinese recognition showed higher averaged classification accuracies and smaller standard deviations across all the sensor groups when compared with the English tasks, especially for the F-40 and NC-40 groups.



Symbols	F-40	NC-40	NO-40	NE-40	NA-80
Groups	Facial	Neck_Central	Neck_Odd	Neck_Even	Neck_All
Channel numbers	40	40	40	40	80
Mean (%)	77.92	78.16	87.75	89.74	94.63
Std (%)	10.03	10.53	6.06	6.04	3.84

(a) English

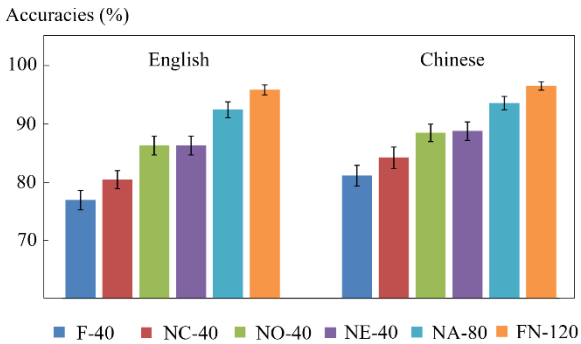


Symbols	F-40	NC-40	NO-40	NE-40	NA-80
Groups	Facial	Neck_Central	Neck_Odd	Neck_Even	Neck_All
Channel numbers	40	40	40	40	80
Mean (%)	85.12	83.02	89.58	91.58	95.09
Std (%)	6.53	8.25	5.71	4.89	3.18

(b) Chinese

Fig. 7. The comparison of the classification accuracies (averaged across different speaking tasks) among different sensor groups for English (a) and Chinese (b) sEMG-based speech recognition.

To further investigate the contributions of different regions of muscles, the classification accuracies averaged across all the different digits and subjects were compared among all the six different sensor groups (F-40, NC-40, NO-40, NE-40, NA-80, and FN-120) for both English and Chinese, as shown in Fig. 8. It was observed that the classification accuracy of the F-40 group was the lowest for both English (76.9%) and Chinese (81.11%) recognition, with the NC-40 group having slightly better performance. The NO-40 and NE-40 groups showed considerably higher accuracies than the F-40 and NC-40 groups, and there were no significant differences between the NO-40 and NE-40 groups. Further increase in the sensor number would also lead to additional performance improvement in the speech classification, such as the NA-80 and FN-120 groups, with the highest accuracy up to 96.54%. It was also observed that the accuracies for English recognition were slightly lower than that of Chinese recognition across all the sensor groups, especially for the F-40 group.



Symbols	F-40	NC-40	NO-40	NE-40	NA-80	FN-120
Groups	Facial	Neck_Central	Neck_Odd	Neck_Even	Neck_All	Face&Neck
Channel numbers	40	40	40	40	80	120
English (Mean ± SE) (%)	76.9±1.67	80.42±1.56	86.28±1.61	86.29±1.62	92.44±1.37	95.85±0.87
Chinese (Mean ± SE) (%)	81.11±1.78	84.21±1.83	88.47±1.50	88.77±1.59	93.58±1.17	96.54±0.73

Fig. 8. The comparison of the classification accuracy averaged across all the different digits and subjects among six different sensor groups for recognizing English and Chinese speaking tasks.

D. Distribution of Optimal Sensors for Different Classification Accuracies

To further localize the best subset of all the HD sEMG sensors that contributed mostly to speech recognition and to reduce the sensor number for practical sEMG-based applications, the SFS algorithm was proposed to automatically find the optimal channel after searching all the 120 sEMG sensors. Then the number of the optimal channels that came from three different sensor groups (F-40, NO-40, and NE-40) was counted for each sensor group, respectively, and the distribution of the optimal channels among the three sensors groups was illustrated in Fig. 9. As shown in Fig. 9(a), as the optimal channel number increased from 5 to 30, the corresponding classification accuracy improved from 74.11% to 94.9% for the English recognition tasks. It was also observed that the optimal channel numbers selected from facial muscles were much less than that from the neck muscles. For instance, for an optimal channel of 5, there was only one optimal channel selected from the F-40 group, while there were both two channels selected from the NO-40 and NE-40 groups. Notably, when the optimal channel number increased, significantly more optimal channels came from the neck region (either the NO-40 or the NE-40 group) instead of the face region. Similar patterns of the optimal channel distribution were also observed for Chinese recognition tasks in Fig. 9(b), with significantly more channels coming from the neck muscles. Moreover, the Chinese recognition tasks seemed to have slightly more optimal sensors coming from the facial muscles (F-40 group), when compared with the English speech recognition.

To further investigate the contributions between the facial and neck muscles for speech recognition, the number of the optimally selected sEMG sensors by the SFS algorithm were statistically analyzed across all the enrolled subjects, and the averaged optimal channel numbers coming from the three different sensor groups (F-40, NO-40, and NE-40) were shown in Fig. 10. It was observed from Fig. 10(a) that classification

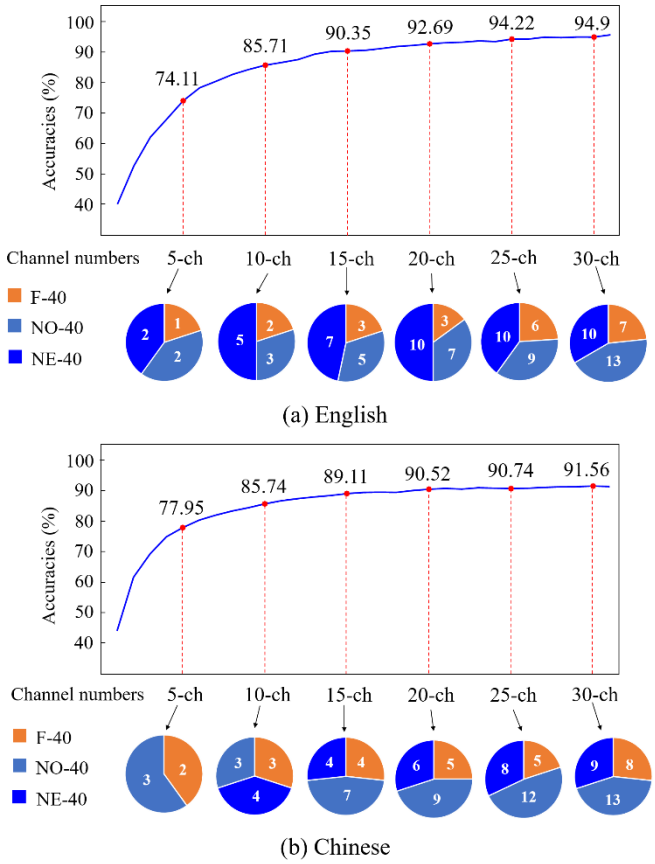


Fig. 9. The distribution of the optimal channels among the three-sensor groups (F-40, NO-40, and NE-40) as a function of the optimal channel number for English (a) and Chinese (b) speech recognitions.

performance showed substantial improvement (from 64.66% to 90.85%) when the optimal channel number increased from 5 to 30, for English speech recognition. The optimal channel number from the F-40 group was significantly lower than either the NO-40 or the NE-40 group, and there was no significant difference between the two groups of the neck region. Similar observations were found for the distribution pattern of the optimal channels for Chinese speech recognition in Fig. 10 (b). It was noteworthy that the average classification accuracies for Chinese recognition were systemically higher than that of English recognition for the same optimally selected channel number.

E. Contributions of Different Muscle Groups With the Increasing Class Number

In addition, to further examine the overall performance of our system, we increased the word number by combining the ten English and ten Chinese speaking tasks as a new set with a total of 20 speech tasks, and then the classification accuracies were compared across different sensor groups (F-40, NE-40, and NO-40) as the class number increased from 1 to 20. As it was shown in Fig. 11 below, the classification accuracies of the F-40 group remained at the lowest level when compared with the NO-40 and NE-40 groups. Meanwhile, a decrease in the classification accuracies was constantly observed when increasing the speech class number, regardless of the sensor

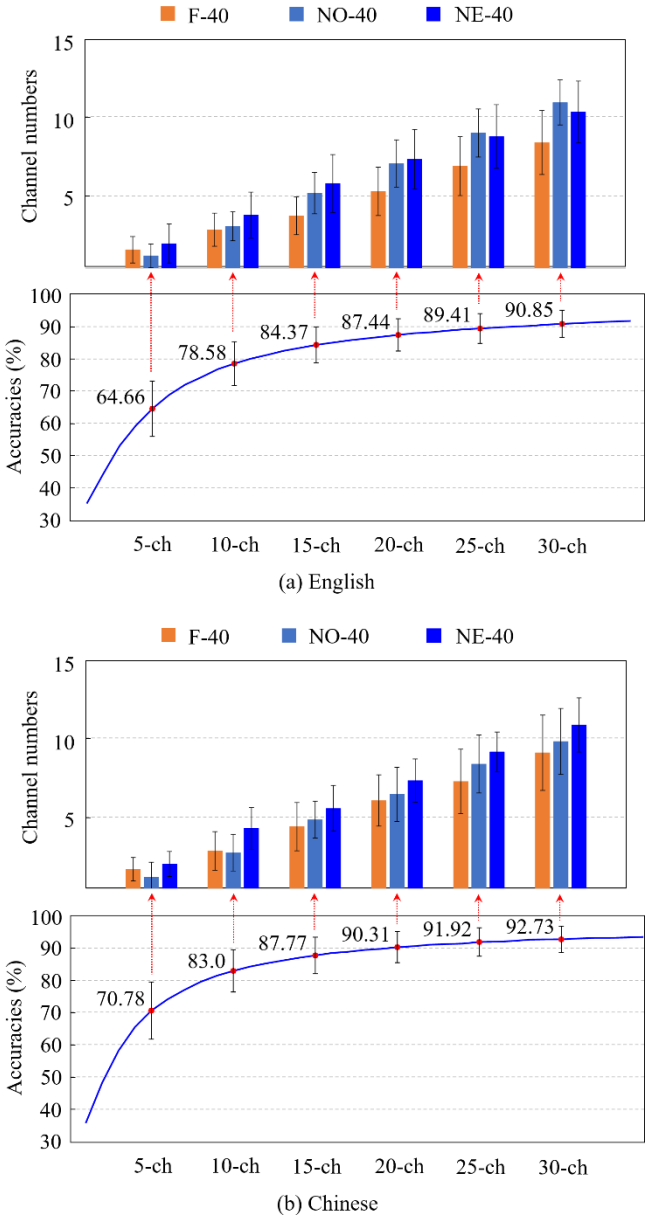


Fig. 10. The distribution of the optimal channel numbers coming from the three different sensor groups (F-40, NO-40, and NE-40) when averaged across all the recruited subjects for English (a) and Chinese (b) speech recognition.

groups (F-40, NO-40, or NE-40). However, the declining rate of accuracy was quite different among different sensor groups. When the class number increased from 1 to 20, the classification accuracies dropped from 100% to 73.76% for the F-40 group, 85.1% for the NO-40 group, and 87.67% for the NE-40 group, respectively. In comparison, the classification accuracies for NO-40 and NE-40 groups were significantly higher than the F-40 group when recognizing 20 words, and the results were consistent with the findings from classifying 10 English or Chinese words.

Moreover, the distribution of the optimally selected sensors was also calculated and compared among the F-40, NO-40, and NE-40 sensor groups when recognizing 20 speaking classes, as shown in Fig. 12. As could be observed from the figure, when the optimal channel number increased from

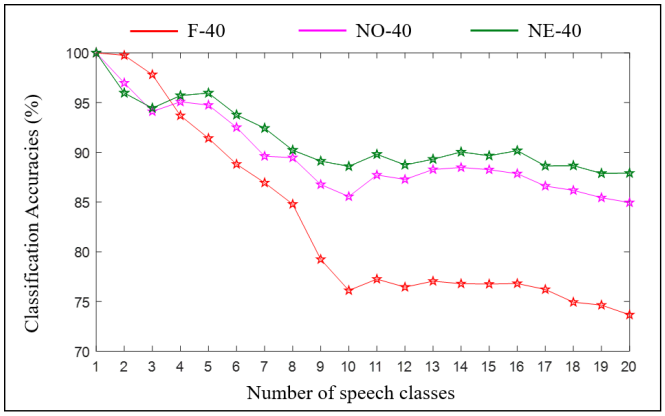


Fig. 11. Comparison of the classification accuracies among the F-40, NO-40, and NE-40 sensor groups when increasing the speech class number from one to twenty.

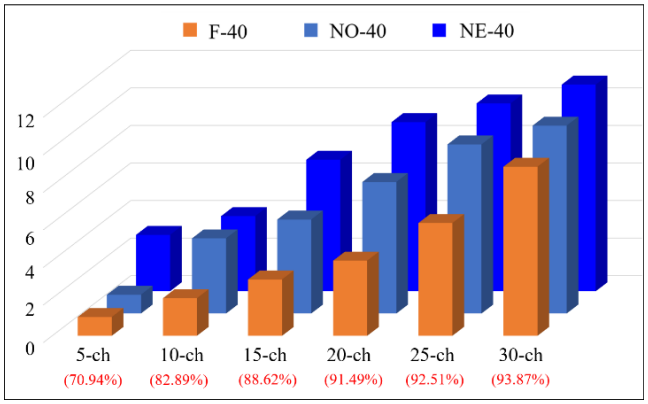


Fig. 12. The distribution of the optimally selected sensors among different sensor groups (F-40, NO-40, and NE-40) when recognizing 20 speaking classes.

5 to 30, the corresponding classification accuracy dramatically increased from 70.94% to 93.87% for recognizing 20 speech classes. By further examining the sensor distribution, it was found that the optimally selected sensors are mainly distributed on the neck muscles (either the NO-40 or the NE-40 group) instead of the face muscles (F-40). For example, among the 5 optimal selected channels, there was only one sensor coming from the F-40 group. When the optimal channel number increased, the number of optimal sensors from the F-40 group was always smaller than the NO-40 or NE-40 group, similar to the findings of 10 English or Chinese word classes.

IV. Discussion

The sEMG-based ASR is a technique that enables the recognition of speaking activities into a textual representation using the sEMG signals recorded from the articulatory muscles associated with speaking activities by the sEMG sensors. The principal objective of this study was to examine the contributions of different articulatory muscles for the sEMG-based ASR, which would be helpful for providing practical guidelines for sEMG sensor placement. This purpose was achieved by using the HD sEMG signals recorded from the facial and neck muscles when speaking ten digits in English and Chinese, respectively.

The study showed that the energy maps calculated from the HD sEMG signals could help to visualize the dynamic energy distribution of the muscular activities during the speaking process (Fig. 4) and provide physiological clues to identify different word pronunciations (Fig. 5). The HD sEMG topographic energy maps are attributed to the vocal cord vibration and mouth movement during the physiological process of speaking [52], [53]. The dynamic spatiotemporal patterns in normal subjects (Fig. 4 and 5) could illustrate the characteristics of a normal speaking process and, therefore, could establish a standard for the diagnosis of the articulatory muscle activities. Meanwhile, the placement of the electrode used for evaluating the speaking functions should follow the myoelectrical characteristics of the speaking activities. Based on the results of this study, the electrodes located in the center of the neck or close to the mouth picked up the largest amplitude of sEMG signals, and therefore they are important for providing the most reliable information for speaking assessment. The findings would suggest that the HD sEMG topographic energy maps could be possibly used as a potential tool for finding the proper sensor placement for speaking related researches, such as speech recognition or evaluation of phonation function. It should also be noted that there were significant individual differences in the classification accuracies when using the same group of sensors (Fig. 7), which might be a result of the different speaking styles or habits of different individuals or languages. Therefore, the purpose of this study is to obtain an individual-independent general understanding of the contributions of different articulatory muscles and therefore provide practical guidelines for sensor placements that are applicable to all individuals.

The use of the HD sEMG technique with multi-channel sensors plays an important role in the investigations of this study by means of covering all the small articulatory muscles in high space-resolution and providing full information about the muscular activities during the speaking process. In most of the previous studies, the sEMG-based ASR investigations depended on the sEMG signals recorded from a few numbers of sensors whose positions were chosen empirically with no quantitative analysis, such as five facial sensors for Thai word recognition [11], eight face and neck sensors for English silent speech recognition [10], and ten face and neck surface sensors for Chinese silent speech classification [13]. However, the insufficient small number of sEMG sensors chosen by experience might lead to the missing of important muscle coverage and major electrical activities that would be essential for speech recognition. For example, placing all the sEMG sensors along the edges of the face or the neck region (Fig. 1) may miss the large amplitude of muscular activities in the middle regions and result in the rather low amplitude of sEMG signals containing little information about the speaking activities (Fig. 4 and 5), leading to the deterioration of the classification performance of the speech recognition. The HD sEMG technique utilized a total of 120 sEMG sensors that are enough to cover all the face and neck muscles, and ensured that no important information about the muscular activities was missed to investigate the contributions of different muscles thoroughly.

In this study, the 120 HD sEMG sensors were divided into six different groups based on their locations to assess the contributions of different articulatory muscles for the sEMG-based ASR systems. The results from Fig. 7 and 8 showed that the facial sensor group (F-40) had significantly lower classification accuracies than any of the neck sensor groups (NC-40, NO-40, NE-40), although with the same channel numbers. Meanwhile, the results in Fig. 11 showed that the classification accuracy of the facial group F-40 was the lowest compared with the sensor groups on the neck (NO-40 and NE-40), when the class number of speaking tasks increased from one to twenty. These results demonstrated that the neck muscles should be the main contributor towards satisfactory speech recognition performance. The findings confirm that the placement of the sensors greatly affects the classification rate of the ASR system, and the neck muscles show a more important role in better speech recognition than the facial muscles. It may be explained by the physiological fact that there are more articulatory muscles distributed along the neck regions, and there are more muscles activated or involved during speech production [27]. The insignificant differences between the NO-40 and NE-40 groups may be attributed to the reason that the two groups are interlaced with extremely close space between neighboring columns to cover nearly the same information sources. The finding that the NC-40 group had significantly lower classification accuracy than either the NO-40 or the NE-40 group suggests that the sEMG sensors should cover larger areas to achieve better classification performances, which is also consistent with the findings of our previous studies on sEMG-based speech recognition [54], [55]. These findings of this study may be useful for providing useful recommendations about sensor placements in routine practices of sEMG-based speech recognition.

Considering that there could be redundancy within the HD sEMG signals and placing sEMG sensors as many as 120 could be time-consuming, the SFS algorithm was proposed to automatically select the optimal channels with the highest classification accuracy so that the sensor number could be greatly reduced. The results in Fig. 9 showed that the classification accuracy dramatically increased with the optimal channel number, and it could reach about 90% for only 15 optimally selected sensors. By further analyzing the origination, it was found that significantly more optimal sensors came from the neck sensor group (either NO-40 or NE-40) than the facial sensor group (F-40), although the sensor groups had the same number of channels. Besides, similar findings were shown in Fig. 12, the classification accuracy was 88.62% with only 15 optimally selected sensors when mixing all the English and Chinese words. These findings indicated that the neck muscles should be a more significant contributor to sEMG-based speech recognition, which agrees with the findings in Fig. 7 and 8. It may be explained by the fact that speeches were generated by the quasi-periodic vibration of the vocal cords located within the larynx, which were mainly controlled by the articulatory muscles around the neck. The results of our study suggest that instead of placing an equal number of sensors on the face and neck, it may be a better practice to place more sensors along the

neck region to further improve the classification performance. Other approaches besides the SFS algorithm could also be used to further reduce the number of the sensor channels, so the wearable sEMG-based speech recognition systems or devices could be developed by placing only a few electrodes on the optimal locations. Acoustic and inertial sensors could also be employed in future studies so that the information from different types of sensors could be fused to reduce more channels and additionally improve the performance of the sEMG-based speech recognition.

The language also acts as an essential factor in the speech recognition, and many different sEMG-based speech recognition systems were developed in previous studies for different languages, such as English [14], [15], Chinese [16], [17], Japanese [20], Portuguese [56], Spanish [22] and Arabic [8]. In this study, the performances of sEMG-based speech recognition were systemically compared between English and Chinese under different conditions. The results of Fig. 8, 9, and 10 indicated that the classification accuracies of Chinese speech recognition were slightly higher than English recognition, regardless of the sensor groups or the optimal sensor channels. The slight superior performance of Chinese speech recognition may be attributed to the fact that all the recruited subjects are native Chinese speakers, and they were more fluent in Chinese speaking. It was also observed that when using only the facial sensors (F-40) or the neck sensors (NO-40 or NE-40), the English recognition showed significantly lower classification accuracies than Chinese, indicating that the English-speaking task may rely more heavily on the coordination between the facial and neck muscles. However, only ten digits were employed in the experiments of this study; more different words or phonemes could be involved for further investigating the differences between English and Chinese speech recognition in future studies.

V. Conclusion

In this study, multi-channel sEMG sensors (120 channels) were placed on the facial and neck muscles with high spatial resolution, and the recorded HD sEMG signals were used for automatic speech recognition of English and Chinese digits. The energy maps calculated from the HD sEMG signals showed that the muscular activities of different locations demonstrated significant patterns during the speaking process, and they could help to visualize the dynamic energy distribution of the articulatory muscular activities. The classification accuracies when using only the sensors on the face were significantly lower than those for the neck muscles, although with the same number of channel numbers. The optimal sensors automatically selected by the sequential forward selection algorithm mainly distributed along with muscles on the neck instead of the face. The classification accuracies of Chinese speech recognition were slightly higher than English recognition, regardless of the sensor groups or the optimal channel number. The findings of this study showed that the multi-channel sEMG sensors could be useful to study the muscular activation patterns during speech recognition comprehensively, and the muscles on the neck should be the main contributor towards satisfactory classification performance. This study could provide valuable clues for the development of a practical sEMG-based speech recognition system, especially for patients with speaking disorders.

Acknowledgment

The authors would like to thank all the members of our research laboratory at the Research Center for Neural Engineering, Institute of Advanced Integration Technology, Shenzhen Institutes of Advanced Technology, for their supports and assistance in conducting the experiments and signal processing.

Mingxing Zhu, Xiaochen Wang, Xin Wang, Cheng Wang, and Haoshi Zhang are with the CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China, also with the Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen, 518055 China, and also with the Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: mx.zhu@siat.ac.cn; xc.wang@siat.ac.cn; wangxin@siat.ac.cn; cheng.wang2@siat.ac.cn; zhanghaoshi@siat.ac.cn).

Zhen Huang is with the Department of Rehabilitation Medicine, Guangzhou Panyu Central Hospital, Guangzhou 511400, China (e-mail: mishz@126.com).

Guoru Zhao, Shixiong Chen, and Guanglin Li are with the CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China, and also with the Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen 518055, China (e-mail: gr.zhao@siat.ac.cn; sx.chen@siat.ac.cn; gl.li@siat.ac.cn).

References

[1] T. Kubo, M. Yoshida, T. Hattori, and K. Ikeda, "Towards excluding redundancy in electrode grid for automatic speech recognition based on surface EMG," *Neurocomputing*, vol. 134, pp. 15–19, Jun. 2014.

[2] A. Rameau, "Pilot study for a novel and personalized voice restoration device for patients with laryngectomy," *Head Neck*, vol. 42, no. 5, pp. 839–845, May 2020.

[3] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline, "Silent speech recognition as an alternative communication device for persons with laryngectomy," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 12, pp. 2386–2398, Dec. 2017.

[4] Q. Ai, W. Zhang, B. Zhang, G. Li, and M. Yang, "Convolutional neural network applied in mime speech recognition using sEMG data," in *Proc. Chin. Autom. Congr. (CAC)*, Hangzhou, China, Nov. 2019, pp. 3347–3352.

[5] Y. Deng, G. Colby, J. T. Heaton, and G. S. Meltzner, "Signal processing advances for the MUTE sEMG-based silent speech recognition system," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Orlando, FL, USA, Oct. 2012, pp. 1–6.

[6] P. Das, B. Neogi, A. Chandra, and A. Dey, "Unvoiced speech recognition using dynamic analysis of EMG signal," in *Computational Advancement in Communication Circuits and Systems* (Lecture Notes in Electrical Engineering). Cham, Switzerland: Springer, Jan. 2020, pp. 217–225.

[7] B. J. Betts, K. Binsted, and C. Jorgensen, "Small-vocabulary speech recognition using surface electromyography," *Interacting Comput.*, vol. 18, no. 6, pp. 1242–1259, Dec. 2006.

[8] L. Fraiwan, K. Lweesy, A. Al-Nemrawi, S. Addabass, and R. Saifan, "Voiceless arabic vowels recognition using facial EMG," *Med. Biol. Eng. Comput.*, vol. 49, no. 7, pp. 811–818, Mar. 2011.

[9] S. S. Mostafa, M. A. Awal, M. Ahmad, and M. A. Rashid, "Voiceless Bangla vowel recognition using sEMG signal," *SpringerPlus*, vol. 5, no. 1, p. 1522, Dec. 2016.

[10] N. Srisuwan, P. Phukpattaranont, and C. Limsakul, "Comparison of feature evaluation criteria for speech recognition based on electromyography," *Med. Biol. Eng. Comput.*, vol. 56, no. 6, pp. 1041–1051, Nov. 2017.

[11] N. Sae Jong and P. Phukpattaranont, "A speech recognition system based on electromyography for the rehabilitation of dysarthric patients: A thai syllable study," *Biocybernetics Biomed. Eng.*, vol. 39, no. 1, pp. 234–245, Jan. 2019.

[12] M. Zhang, Y. Wang, Z. Wei, M. Yang, Z. Luo, and G. Li, "Inductive conformal prediction for silent speech recognition," *J. Neural Eng.*, pp. 1–12, Mar. 2020, doi: 10.1088/1741-2552/ab7ba0.

[13] G. S. Meltzner, G. Colby, Y. Deng, and J. T. Heaton, "Signal acquisition and processing techniques for sEMG based silent speech recognition," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2011, pp. 4848–4851.

[14] L. Diener, S. Bredehoeft, and T. Schultz, "A comparison of EMG-to-Speech Conversion for Isolated and Continuous Speech," in *Speech Communication; 13th ITG-Symposium*, vol. 2018, pp. 1–5.

[15] M. Janke and L. Diener, "EMG-to-speech: Direct generation of speech from facial electromyographic signals," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 12, pp. 2375–2385, Dec. 2017.

[16] X. Yang, X. Chen, X. Cao, S. Wei, and X. Zhang, "Chinese sign language recognition based on an optimized tree-structure framework," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 4, pp. 994–1004, Jul. 2017.

[17] Y. Li, X. Chen, X. Zhang, K. Wang, and Z. J. Wang, "A sign-component-based framework for Chinese sign language recognition using accelerometer and sEMG data," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 10, pp. 2695–2704, Oct. 2012.

[18] J. Cheng, X. Chen, A. Liu, and H. Peng, "A novel phonology- and radical-coded chinese sign language recognition framework using accelerometer and surface electromyography sensors," *Sensors*, vol. 15, no. 9, pp. 23303–23324, Sep. 2015.

[19] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Commun.*, vol. 52, no. 4, pp. 270–287, Apr. 2010.

[20] T. Kubo, T. Toda, M. Yoshida, T. Hattori, and K. Ikeda, "Vowel recognition based on surface electromyography with electrode grid on submental region," *Trans. JSME*, vol. 50, no. 1, pp. 38–46, Jan. 2012.

[21] N. Srisuwan, P. Phukpattaranont, and C. Limsakul, "Feature selection for thai tone classification based on surface EMG," *Procedia Eng.*, vol. 32, pp. 253–259, dec2012.

[22] N. Srisuwan, P. Prukpattaranont, and C. Limsakul, "Comparison of classifiers for EMG based speech recognition," *J. Phys., Conf. Ser.*, vol. 1438, Jan. 2020, Art. no. 012032.

[23] K.-S. Lee, "EMG-based speech recognition using hidden Markov models with global control variables," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 3, pp. 930–940, Mar. 2008.

[24] D. S. Putra and Y. U. Ww, "Feature extraction of facial electromyograph (EMG) signal for aceh languages speech using discrete wavelet transform (DWT)," *J. Inotera*, vol. 4, no. 1, pp. 31–40, Jun. 2019.

[25] R. Abdullah, H. Muthusamy, V. Vijean, Z. Abdullah, and F. N. C. Kassim, "Real and complex wavelet transform approaches for malaysian speaker and accent recognition," *Pertanika. J. Sci. Technol.*, vol. 27, no. 2, pp. 737–752, Apr. 2019.

[26] P. Gómez-Vilda *et al.*, "Neuromechanical modelling of articulatory movements from surface electromyography and speech formants," *Int. J. Neural Syst.*, vol. 29, no. 2, Mar. 2019, Art. no. 1850039.

[27] J. Apps, *Voice and Speaking Skills for Dummies*. Hoboken, NJ, USA: Wiley, 2012, pp. 82–99.

[28] T. Scott-Phillips, *Speaking Our Minds: Why Human Communication Is Different, and How Language Evolved to Make it Special* (Macmillan International Higher Education). London, U.K.: Red Globe Press, 2014, pp. 124–126.

[29] X. Wu, J. Dang, and I. Stavness, "Iterative method to estimate muscle activation with a physiological articulatory model," *Acoust. Sci. Technol.*, vol. 35, no. 4, pp. 201–212, 2014.

[30] A. J. Young, L. J. Hargrove, and T. A. Kuiken, "The effects of electrode size and orientation on the sensitivity of myoelectric pattern recognition systems to electrode shift," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 9, pp. 2537–2544, Sep. 2011.

[31] H. Nakamura, Y. Konishi, and M. Yoshida, "High-density EMG techniques in neuromuscular studies," *Trans. JSME*, vol. 51, p. M-17, Mar. 2013.

[32] B. Afsharipour, K. Ullah, and R. Merletti, "Amplitude indicators and spatial aliasing in high density surface electromyography recordings," *Biomed. Signal Process. Control*, vol. 22, pp. 170–179, Sep. 2015.

[33] D. Wang, X. Zhang, X. Gao, X. Chen, and P. Zhou, "Wavelet packet feature assessment for high-density myoelectric pattern recognition and channel selection toward stroke rehabilitation," *Frontiers Neurol.*, vol. 7, no. 3, p. 197, Nov. 2016.

[34] D. Bai, S. Chen, and J. Yang, "Upper arm motion high-density sEMG recognition optimization based on spatial and time-frequency domain features," *J. Healthcare Eng.*, vol. 2019, pp. 1–16, Mar. 2019.

[35] Stegeman, F. Dick, B. U. Kleine, B. G. Lapatki, and J. P. Van Dijk, "High-density surface EMG: Techniques and applications at a motor unit level," *Biocybernetics Biomed. Eng.*, vol. 32, no. 3, pp. 3–27, Jan. 2012.

[36] T. Sa-Ngiamsak, J. Costa, and J. Baptista, "High-density surface electromyography applications & reliability vs muscle fatigue—A short review," in *Proc. Occupational Safe. Hygiene II*, Jan. 2014, pp. 265–369.

[37] A. Phinyomark, P. Phukpattaranont, and C. Limsakul, "Feature reduction and selection for EMG signal classification," *Expert Syst. Appl.*, vol. 39, no. 8, pp. 7420–7431, Jun. 2012.

[38] M. Hakonen, H. Piitulainen, and A. Visala, "Current state of digital signal processing in myoelectric interfaces and related applications," *Biomed. Signal Process. Control*, vol. 18, pp. 334–359, Apr. 2015.

[39] Y.-C. Du, C.-H. Lin, L.-Y. Shyu, and T. Chen, "Portable hand motion classifier for multi-channel surface electromyography recognition using grey relational analysis," *Expert Syst. Appl.*, vol. 37, no. 6, pp. 4283–4291, Jun. 2010.

[40] K. S. Kim, H. H. Choi, C. S. Moon, and C. W. Mun, "Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions," *Current Appl. Phys.*, vol. 11, no. 3, pp. 740–745, May 2011.

[41] F. D. Farfán, J. C. Politti, and C. J. Felice, "Evaluation of EMG processing techniques using information theory," *Biomed. Eng. OnLine*, vol. 9, no. 1, p. 72, 2010.

[42] G. Li, J. Li, Z. Ju, Y. Sun, and J. Kong, "A novel feature extraction method for machine learning based on surface electromyography from healthy brain," *Neural Comput. Appl.*, vol. 31, no. 12, pp. 9013–9022, Mar. 2019.

[43] A. Phinyomark, F. Quaine, S. Charbonnier, C. Serviere, F. Tarpin-Bernard, and Y. Laurillau, "A feasibility study on the use of anthropometric variables to make muscle–computer interface more practical," *Eng. Appl. Artif. Intell.*, vol. 26, no. 7, pp. 1681–1688, Aug. 2013.

[44] J. Liu, "Adaptive myoelectric pattern recognition toward improved multifunctional prosthesis control," *Med. Eng. Phys.*, vol. 37, no. 4, pp. 424–430, Apr. 2015.

[45] F. Palermo, M. Cognolato, A. Gijsberts, H. Müller, B. Caputo, and M. Atzori, "Repeatability of grasp recognition for robotic hand prosthesis control based on sEMG data," in *Proc. Int. Conf. Rehabil. Robot. (ICORR)*, Jul. 2017, pp. 1154–1159.

[46] W.-T. Shi, Z.-J. Lyu, S.-T. Tang, T.-L. Chia, and C.-Y. Yang, "A bionic hand controlled by hand gesture recognition based on surface EMG signals: A preliminary study," *Biocybernetics Biomed. Eng.*, vol. 38, no. 1, pp. 126–135, 2018.

[47] M. Jochumsen, A. Waris, and E. N. Kamavuako, "The effect of arm position on classification of hand gestures with intramuscular EMG," *Biomed. Signal Process. Control*, vol. 43, pp. 1–8, May 2018.

[48] O. W. Samuel, Y. Geng, X. Li, and G. Li, "Towards efficient decoding of multiple classes of motor imagery limb movements based on EEG spectral and time domain descriptors," *J. Med. Syst.*, vol. 41, no. 12, p. 194, Oct. 2017.

[49] O. W. Samuel *et al.*, "Pattern recognition of electromyography signals based on novel time domain features for amputees' limb motion classification," *Comput. Electr. Eng.*, vol. 67, pp. 646–655, Apr. 2018.

[50] X. Li, O. W. Samuel, X. Zhang, H. Wang, P. Fang, and G. Li, "A motion-classification strategy based on sEMG-EEG signal combination for upper-limb amputees," *J. NeuroEngineering Rehabil.*, vol. 14, no. 1, Jan. 2017.

[51] D. P. Campos *et al.*, "Single-channel sEMG dictionary learning classification of ingestive behavior on cows," *IEEE Sensors J.*, vol. 20, no. 13, pp. 7199–7207, Jul. 2020.

[52] M. F. Regner, C. Tao, D. Ying, A. Olszewski, Y. Zhang, and J. J. Jiang, "The effect of vocal fold adduction on the acoustic quality of phonation: Ex vivo investigations," *J. Voice*, vol. 26, no. 6, pp. 698–705, Nov. 2012.

[53] R. Mittal, B. D. Erath, and M. W. Plesniak, "Fluid dynamics of human phonation and speech," *Annu. Rev. Fluid Mech.*, vol. 45, no. 1, pp. 437–467, Jan. 2013.

[54] J. Zhuang *et al.*, "Comparison of contributions between facial and neck muscles for speech recognition using high-density surface electromyography," in *Proc. IEEE Int. Conf. Comput. Intell. Virtual Environ. Meas. Syst. Appl. (CIVEMSA)*, Tianjin, China, Jun. 2019, pp. 1–5.

[55] M. Zhu *et al.*, "Comparison of english and Chinese speech recognition using high-density electromyography," in *Proc. 13th Int. Conf. Sens. Technol. (ICST)*, Dec. 2019, pp. 1–5.

[56] J. Freitas, A. Teixeira, and M. S. Dias, "Towards a silent speech interface for Portuguese surface electromyography and the nasality challenge," in *Proc. 5th Int. Conf. Bio. Eng. Technol. (BIOSTEC)*, Vilamoura, Portugal, Jan. 2012, pp. 91–100.

Mingxing Zhu (Graduate Student Member, IEEE) received the master's degree in control science and engineering from Central South University, Changsha, China, in 2012. She is currently pursuing the Ph.D. degree with the Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences. She joined the Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences, Shenzhen, China, in 2012, working as an Engineer in the Research Centre for Neural Engineering, Institute of Integrated Technology. Her research direction is high-density surface electromyograph technique application.

Zhen Huang received the bachelor's degree in traditional Chinese medicine from the Jiangxi University of Traditional Chinese Medicine, China, in 1986. She is currently the Director of the Department of Rehabilitation Medicine, Central Hospital of Panyu District, Guangzhou, and the Deputy Director of the Rehabilitation Research Institute of Panyu District, Guangzhou. Her research direction is neurorehabilitation, cardiopulmonary rehabilitation, pain rehabilitation, and other fields.

Xiaochen Wang (Graduate Student Member, IEEE) received the bachelor's degree in software engineering from Tianjin University of Finance and Economics in 2017, and the master's degree from the Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, China, in 2018. Her research interest includes the acquisition and processing of high-density surface electromyography signals.

Xin Wang (Associate Member, IEEE) received the bachelor's degree in electrical engineering and automation from the Wuhan University of Science and Technology, Wuhan, China, in 2017. He is currently pursuing the master's degree with the Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, China. His research interests include biomedical signal processing and audiometry technology.

Cheng Wang (Graduate Student Member, IEEE) received the bachelor's degree in information management and information system from the Harbin Institute of Technology, China, in 2019. He is currently pursuing the master's degree with the Center for Neuroengineering, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. His research interests include the acquisition and processing of biomedical signals, analyzing brain cognitive impairment activities, and machine learning.

Haoshi Zhang received the master's degree in biomedical engineering from Xi'an Jiaotong University, Xi'an, China, in 2010. She joined the Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences, Shenzhen, China, in 2010, working as an Engineer at the Research Centre for Neural Engineering, Institute of Integrated Technology. Her research direction is biomedical signal processing. She focused on EMG and EEG signal processing.

Guoru Zhao (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from Jilin University, China, in 2003, 2006, and 2009, respectively. He had studied at the RoyalVeterinary College, University of London, U.K., as a Chinese Government Sponsored Scholar from 2007 to 2008. He is currently a Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His current research interests mainly focus on rehabilitation biomechanics and robotic systems, and artificial intelligence algorithm.

Shixiong Chen (Member, IEEE) received the bachelor's and master's degrees in biomedical engineering from Tsinghua University, China, in 2005 and 2007, respectively, and the Ph.D. degree in speech and hearing sciences from Arizona State University, Tempe, AZ, USA, in 2012. He is a Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research interests include the acquisition and processing of biomedical signals (EEG, EMG, and ECG), rehabilitation technologies of hearing disorders, and medical device instrumentation. He is the primary investigator of multiple national/governmental research grants and the Associate Director of the Shenzhen Engineering Laboratory of Neuro-Rehabilitation Technology.

Guanglin Li (Senior Member, IEEE) received the Ph.D. degree in biomedical engineering from Zhejiang University, China, in 1997. From 1999 to 2002, he was a Postdoctoral Research Associate with the Department of Bioengineering, The University of Illinois at Chicago. From 2002 to 2006, he was a Senior Research Scientist with BioTechPlex Corporation. From 2006 to 2009, he served as a Senior Research Scientist with the Neural Engineering Center for Artificial Limbs, Rehabilitation Institute of Chicago, and jointly served as an Assistant Professor of Physical Medicine and Rehabilitation at the Northwestern University. Since 2009, he has been with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, where he is currently a Professor. He has also served as the Director of the CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems since 2014. He has authored over 120 peer-reviewed papers and filed over 50 patents in the field of biomedical engineering and rehabilitation engineering. His current research interests include neuro-rehabilitation engineering, human–machine interaction, rehabilitation robotics, flexible sensing technologies, and neural functional reconstruction.