Decoding Silent Speech Based on High-Density Surface Electromyogram Using Spatiotemporal Neural Network

Xi Chen, *Student Member, IEEE*, Xu Zhang, *Member, IEEE*, Xiang Chen, *Member, IEEE*, and Xun Chen, *Senior Member, IEEE*

***Abstract*— Finer-grained decoding at a phoneme or syllable level is a key technology for continuous recognition of silent speech based on surface electromyogram (sEMG). This paper aims at developing a novel syllable-level decoding method for continuous silent speech recognition (SSR) using spatio-temporal end-to-end neural network. In the proposed method, the high-density sEMG (HD-sEMG) was first converted into a series of feature images, and then a spatio-temporal end-to-end neural network was applied to extract discriminative feature representations and to achieve syllable-level decoding. The effectiveness of the proposed method was verified with HD-sEMG data recorded by four pieces of 64-channel electrode arrays placed over facial and laryngeal muscles of fifteen subjects subvocalizing 33 Chinese phrases consisting of 82 syllables. The proposed method outperformed the benchmark methods by achieving the highest phrase classification accuracy (97.17 ± 1.53%, $p < 0.05$), and lower character error rate (3.11 ± 1.46%, $p < 0.05$). This study provides a promising way of decoding sEMG towards SSR, which has great potential applications in instant communication and remote control.**

***Index Terms*— Silent speech recognition, high-density surface electromyography, spatiotemporal feature, language model, time sequence decoding.**

I. Introduction

SPEECH, is at once the most natural way of communicating with others and also a survival skill. Automatic speech recognition (ASR) is a common speech technology in people's daily life that can recognize the speaker's intention, while its development has greatly promoted the relationship between humans and computers, enabling computers to better understand human language for natural and robust human-computer interaction. Common applications include digital personal care [1], smart home [2], smart medical [3], [4], etc. Although ASR has achieved great success, its use in some special scenarios is still limited. These include the degradation of ASR performance in the presence of acoustic noises and the privacy concern during communication in public places. In addition, people who lost the ability to speak due to laryngeal disease or other reasons cannot benefit from ASR-related technologies. Silent speech recognition (SSR) technology provides a solution to the aforementioned challenges because it does not rely on acoustic signals but other medium. Several signal modalities have been applied to realize SSR by capturing the movement of articulatory muscles or extracting neural information, such as the electromagnetic arthrography [5], the ultrasound or optical images of tongue or lips [6], [7], [8], the electromyogram (EMG) [9], [10], [11], [12], and the electroencephalogram [13], [14].

The surface EMG (sEMG) is an optional choice for SSR, and it could record muscular activities related to vocalization by collecting the electrophysiological signals from the skin surface using non-invasive electrodes [15], [16]. The sEMG signal can be viewed as a command source to decode the neural commands related to movement for establishing a generalized neuro-machine interface towards myoelectric control of orthotic robots, hand prostheses and wearable devices [17], [18], [19]. In the process of speech production, the movement of articulatory muscles, i.e., facial and laryngeal muscles, is the intuitive response to the vocal nerve commands of human body, and the decoding of sEMG signals corresponds to the reconstruction of the speaker's speech intention. Therefore, the sEMG-based SSR is actually a branch of the myoelectric control technology to decode the intention from the movement of facial and laryngeal muscles, which can be seen as a kind of application case.

Predecessors have done a lot of researches on sEMG-based speech recognition and most of them focused on isolated words with pattern classification algorithms. A variety of algorithms including linear discriminant analysis (LDA) [20], [21], support vector machine (SVM) [16], random forest (RF) [22]were employed to build pattern classifiers. In addition, with the development of biosensing technology, high-density (HD) electrode array is widely applied to simultaneously record multichannel sEMG signals from a number of target muscles or muscle groups in relatively large areas, and the use of HD-sEMG signals has been proved to promote the development of myoelectric control system

Manuscript received 30 August 2022; revised 2 February 2023; accepted 1 March 2023. Date of publication 11 April 2023; date of current version 26 April 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62271464. *(Corresponding author: Xu Zhang.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board.

The authors are with the School of Information Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, China (e-mail: xuzhang90@ustc.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2023.3266299

as well [23], [24], [25]. Some researchers applied the HD electrode arrays to the sEMG-based SSR, demonstrating that their spatial information of muscle activities can improve the SSR performance [26], [27], [28]. It was worth note that the HD-sEMG data recorded by two-dimensional electrode arrays can be viewed as an image, where discriminative spatial features can be well characterized by deep image processing techniques such as convolutional neural networks (CNNs) [29]. Nevertheless, these common classification algorithms treated a phrase or a sequence of words as a single pattern, ignoring the grammatical or contextual relation between syllables and phonemes. Continuous SSR capable of decoding semantics has always been a pursuit.

Based on these considerations, some researchers made attempts on continuous speech recognition with prior semantic knowledge. The three-state left-to-right fully continuous hidden Markov models (HMMs) with a Gaussian mixture model emission probability were designed to realize continuous speech recognition, and they demonstrated how to train the phoneme-based acoustic models using sEMG [30]. Similarly, another study reported a method using HMMs to achieve speech recognition at a word and phoneme level using sEMG, achieving a word error rate of 8.90% on a 2200-word vocabulary [9]. In such an HMM-based recognition algorithm, many essential models need to be trained separately, such as the acoustic model, the pronunciation dictionary model, and the language model (LM). These models played different roles in speech recognition tasks and they required not only a large number of time-alignment training materials, but also repeated experiments to determine each model's parameters. These limitations raise the technical threshold of SSR applications, which are not conducive to the popularization of SSR systems.

In recent years, deep learning has been applied to decoding the temporal information and has greatly contributed to the development of speech recognition technology [31], [32], [33], [34], [35]. For example, the bidirectional long short-term memory (BiLSTM) network has a strong advantage in characterizing temporal or sequential information, and it has been widely used in natural language processing [34]. The end-to-end method provides a feasible scheme for speech recognition due to its unique way of handling signal samples and it connects the input end (speech waveform or feature sequence) to the output end (word or syllable sequence) by constructing a neural network. Compared with the conventional ASR systems (i.e., the aforementioned HMM-based method), this end-to-end method assumes all the functions in the neural network, so there is no need for training additional acoustic model, pronunciation dictionary model or LM. Connectionist temporal classification (CTC) is a typical end-to-end model and it maps the original acoustic signal to the sequence of phonemes or syllables without segmenting the training signals in advance, thus removing the need to locate ambiguous label boundaries. Actually, it has been successfully applied in the field of ASR [31], [32], [33]. However, the usability and practicality of these end-to-end methods for sEMG-based speech recognition have not been well investigated.

Inspired by the above considerations, we hypothesized that the end-to-end decoding techniques can significantly improve SSR performance. To evaluate this hypothesis, a novel silent speech decoding method was proposed based on the CTC algorithm, which mainly provided a solution to the time-alignment challenge of continuous silent speech sEMG signals [33], [36]. Besides, the CNN module and the BiLSTM module of deep neural networks were also employed toconstitute a feature extractor, due to their good abilities of characterizing spatial and temporal information from the HD-sEMG data, respectively. The proposed method allows us to decode silent speech continuously at a syllable level, leading to improved performance with a variety of potential applications in instant communication and remote control.

II. Methods

A. Subjects and Experiments

1) Subjects: Fifteen subjects (aged: 21-27 years, mean ± standard deviation: 24.53 ± 1.45 years) without any known of language disability participated in the experiment. All of them are native speakers of Mandarin. This study was approved by the Ethics Review Board of the University of Science and Technology of China (Hefei, Anhui, China). All participants were informed of the experimental procedures in detail, and they signed the informed consent before any experiment.

2) Data Collection Protocols: A total of 64 electrodes were arranged in four pieces of HD electrode arrays, with a bilateral symmetrical design. These arrays were in irregular shapes to better fit uneven surface of the skin as shown in Fig. 2. We made such efforts to ensure a good electrode-skin contact so as to reduce effect of motion artifacts. Two arrays of the face were placed on the buccinators, masseter and orbicularis oris muscle. The arrays placed on the laryngeal muscles were designed to record muscular activities from the cervical muscle and anterior belly of the digastric muscle. For an electrode on each array, there was a round probe with a diameter of 5 mm, constituting one sEMG channel in a monopolar manner with a reference electrode attached to the uricularis posterior behind the ear. The inter-electrode distance between two consecutive electrodes on the laryngeal muscles was 18mm while the distance was 18 mm along the horizontal direction, and 10mm along the vertical direction on the facial muscles, respectively. The ground electrode was attached to the uricularis posterior behind the other ear different from the location of the reference electrode.

During the experiment, the subjects were invited to sit comfortably on a height-adjustable chair. Before attaching the electrode array, 75% medical alcohol was used to clean the surface of the skin. Meanwhile, we applied the double-sided medical-grade adhesive tapes with cutouts to secure the arrays to be placed firmly.

A total of 33 phrases from an 82-character vocabulary made up the corpus and each phrase consisted of a number of 2 to 6 syllables or characters, as shown in Fig. 1. Each character corresponded to a syllable in the Chinese language, so we labeled each syllable by one Chinese character. This set of phrases were determined by referencing common words or

33 phrases											
	Chinese	Pronunciation in Chinese Pinyin	English		Chinese	Pronunciation in Chinese Pinyin	English		Chinese	Pronunciation in Chinese Pinyin	English
T1	前进	[qian'jin]	Forward	T12	初始化	[chu'shi'hua]	Initialization	T23	切换功能	[qie'huan'gong'neng]	Switch function
T2	后退	[hou'tui]	Backward	T13	延时	[yan'shi]	Delay	T24	建立坐标系	[jian'li'zuo'biao'xi]	Axes
T3	左转	[zuo'zhuan]	Turn left	T14	计数	[ji'shu]	Count	T25	撤离	[che'li]	Evacuation
T4	右转	[you'zhuan]	Turn right	T15	旋转	[xuan'zhuan]	Rotation	T26	卧倒	[wo'dao]	Lie down
T5	加速	[jia'su]	Speed up	T16	抓取	[zhua'qu]	Grasp	T27	危险	[wei'xian]	Danger
T6	减速	[jian'su]	Slow down	T17	放置	[fang'zhi]	Place	T28	供水	[gong'shui]	Water supply
T7	上升	[shang'sheng]	Go up	T18	定位原点	[ding'wei'yuan'dian]	Locate the origin	T29	伸长水带	[shen'chang'shui'dai]	Elongate hose
T8	悬停	[xuan'ting]	Hover	T19	寻找目标	[xun'zhao'mu'biao]	Search target	T30	空气耗尽	[kong'qi'hao'jin]	Run out of air
T9	下降	[xia'jiang]	Go down	T20	获取坐标	[huo'qu'zuo'biao]	Get coordinates	T31	请求支援	[qing'qiu'zhi'yuan]	Request assistance
T10	开机	[kai'ji]	Start up	T21	目标检测	[mu'biao'jian'ce]	Target detection	T32	呼吸机故障	[hu'xi'ji'gu'zhang]	Respirator malfunction
T11	关机	[guan'ji]	Power off	T22	处理数据	[chu'li'shu'ju]	Processing data	T33	发现被困人员	[fa'xian'bei'kun'ren'yuan]	Find trapped person

83 basic syllables																																							
0	上	4	位	8	关	12	前	16	升	20	发	24	吸	28	坐	32	寻	36	延	40	找	44	援	48	故	52	机	56	水	60	现	64	空	68	耗	72	计	73	退	80	险
1	下	5	供	9	减	13	功	17	卧	21	取	25	员	29	处	33	尽	37	建	41	抓	45	撤	49	数	53	标	57	求	61	理	65	立	69	能	73	请	77	速	81	障
2	人	6	倒	10	切	14	加	18	危	22	右	26	呼	30	始	34	左	38	开	42	换	46	支	50	旋	54	检	58	测	62	目	66	系	70	获	74	转	78	长	82	-
3	伸	7	停	11	初	15	化	19	原	23	后	27	困	31	定	35	带	39	悬	43	据	47	放	51	时	55	气	59	点	63	离	67	置	71	被	75	进	79	降		

Fig. 1. List of the 33 phrases and 83 basic Chinese syllables.

phrases used in previous studies [20], [21], [37] and expanding them with some useful phrases, to meet the requirements of different interactive applications in silent speech scenarios. Specifically, T1-T9 were related to the spatial motion of the object, such as the unmanned aerial vehicle and unmanned ground vehicle. T10-T24 were closely bound up with industrial control for joystick or mechanical arm with multiple degrees of freedom. T25-T33 were designed from fire-fighting terms to provide clear and concise communication in a noisy fire scene environment. For each phrase, the speakers were asked to read the phrase silently under normal and natural conditions following the voice guide of the computer, in 20 repetitions. A 4-s delay was applied between repetitions to prevent mental or muscular fatigue. These 33 phrases consisted of 82 basic Chinese syllables, labeled as 0-81 in total. Moreover, a "blank" character was marked as 82 with the symbol "_" representing no meaningful output at a particular frame. These characters were regarded as the basic elements in syllable-level decoding.

During the experiment, the sEMG signals first passed through a two-stage amplifier with a gain of 64dB and a band-pass filter of 20-500Hz. Subsequently, these analog sEMG signals were converted to digital signals by the analog-to-digital converter with 1KHz sampling rate. Then the signals were transmitted to a laptop computer via a USB cable for data monitoring and storage.

3) HD-sEMG Images Splicing and Feature Extraction: In order to extract spatial information conveniently from the HD electrode array, we rearranged the position of 64 electrodes into a regular 8 × 8 shape, like an image as shown in Fig. 2, and each pixel of the image corresponds to a sEMG channel. During the process of a phrase phonation, a series of sEMG burst activities can be observed with large-amplitude fluctuations. A routine sEMG amplitude-thresholding algorithm was adopted,with the threshold set to three times as large as the baseline amplitude from 64 channels, to judge the onset and offset of the sEMG activities. Thus, a data segment corresponding to each phrase was determined as a basic sample between both the onset and the offset.

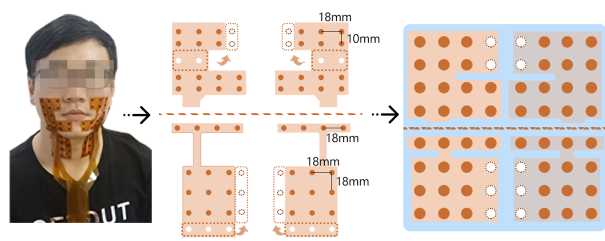

Fig. 2. The display of sEMG electrode positions and scheme for electrode position rearrangement as a regular image.

For each phrase sample, the data segment was further divided into a series of frames. The frame settings needed to be appropriately set to characterize the phoneme/syllable information finely. We conducted sensitivity analyses on both frame length and frame increment, to assess their effects on the SSR performance. The frame length varied from 120ms to 240ms and the frame increment was assigned from 100ms to 200ms, both for every 20 ms. Subsequently, features were extracted for each frame over all channels. A total of four features including mean absolute value of the sEMG signal amplitude and three time-dependent power spectrum descriptors (TD-PSDs) [38], [39] were extracted from each channel of one frame. Thus, we finally get a feature map with the shape of T × 8 × 8 × 4 from a 64-channel sEMG sample, where T was the number of frames of a sample.

B. Spatio-Temporal End-To-End Neural Networks for Syllable-Level Speech Decoding

The neural networks employed in this study consisted of three major modules: a spatial block (equivalent to the

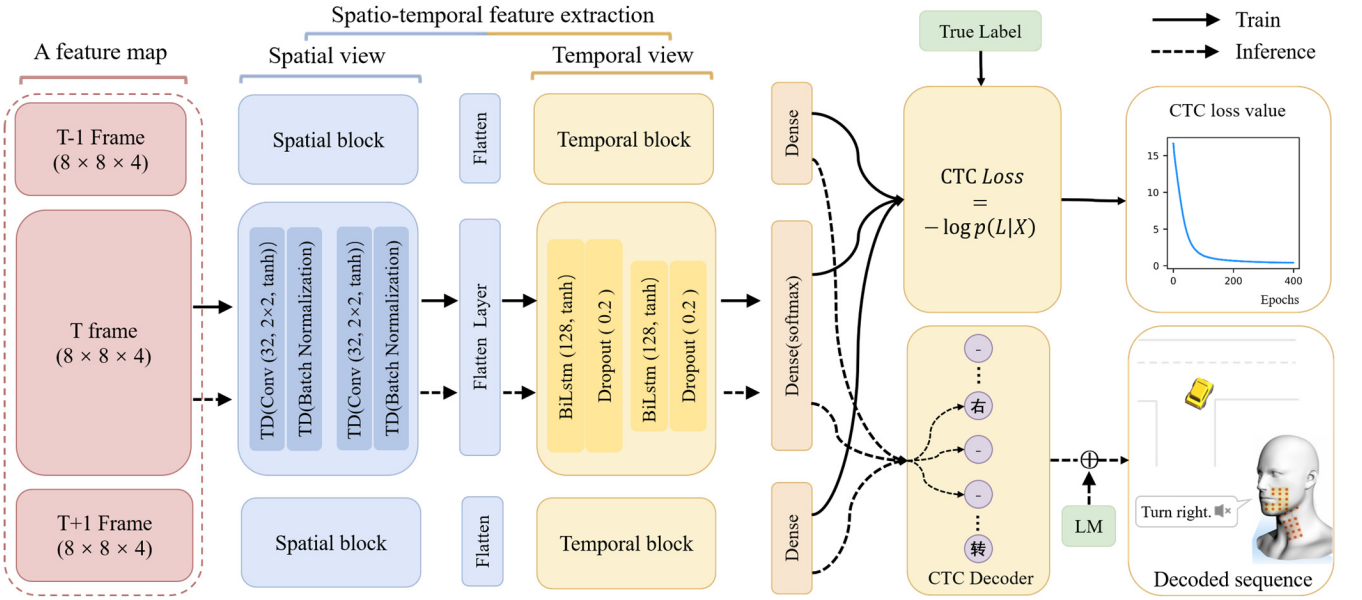

Fig. 3. The architecture of the proposed method.

common CNN module), a temporal block (equivalent to the BiLSTM module) and a CTC decoder, as shown in the Fig. 3. Both the spatial and temporal blocks were designed to well characterize spatiotemporal information. The spatial block consisted of two convolutional layers, each followed by a batch normalization (BN) layer and dropout layer. The BN layer was designed to alleviate this internal covariate shift by introducing a normalization step that fixed the means and variances of layer inputs [40]. The dropout layer was applied after each BN layer with a rate of 20% to avoid overfitting. The feature map was first processed by the spatial block to extract spatial information which represented the activation and location of facial and laryngeal muscles. Next, the spatial block produced these feature representations to the temporal block through the flatten layer to further learn semantic and contextual information. The temporal block comprised two BiLSTM layers, each followed by a dropout layer. Then, the extracted spatio-temporal features were sent to the dense layer for syllable classification with SoftMax activation function and the neural numbers in this dense layer were equal to the number of basic Chinese syllables (that is 83 in this paper). After that, the classification probability matrix (CPM) produced from the dense layer was sent to the CTC decoder for further processing. The CPM showed the probabilities of all syllables to align true sequences with the input sequence. The function of the CTC decoder aims at utilizing the syllable probability to calculate the sequence probability sum of all possible combinations of a true sequence in the training process while searching the sequence related to the maximum decoded sequence probability in the inference process. For a sequence $X = [x_1, x_2, \ldots x_t, \ldots, x_T]$, a probability of sequence π can be calculated as:

$$p(\pi \mid X) = \prod_{t=1}^{T} y_{\pi_t} \tag{1}$$

where π_t represents a syllable at t position of the sequence π and y_{π_t} is the probability of observing the syllable π_t. In this mapping process, every input frame x_t is mapped to a certain label π_t. Any output sequence π generated from the aforementioned last dense layer could be mapped into a target sequence L using the many-to-one mapping function (π). Such function could merge any repetition in a sequence of consecutive non-blank syllables to a single syllable and subsequently remove all blank symbols. Thus, the probability of the target label L is formulated as

$$p(L \mid X) = \sum\nolimits_{\pi \in \beta^{-1}(L)} p(\pi \mid X). \tag{2}$$

The negative log probability was regarded as the loss function for training the neural networks of the proposed method.

$$\mathrm{CTC}Loss = -\log p(L \mid X) \tag{3}$$

A dynamic programming approach was applied to calculate the loss value more efficiently [33].

Given the well-trained networks, any input data sample in the testing phase can finally be inferred into a decoded sequence of syllable decisions by either a beam search or greedy search algorithm [33]. Both were expected to achieve comparable performance, and the proposed method could work by choosing either one. There might be syllable-level errors in the decoded sequence. Such errors could be further corrected due to a limited number of phrases in the corpus. Thus, we adopted a LM using editing distance algorithm [41] to obtain the similarity of the decoded sequence with that of every true phrase, and the phrase with the minimum distance was determined as the final sequence. The training and inference procedures of the neural networks incorporated with the LM are summarized in Algorithm 1.

The network structures and settings were empirically determined by a variety of pretests with the aim of the optimal performance. For training the networks, the Nadam algorithm

Algorithm 1 Training and inference procedure for the proposed method

Input: a feature map (T × 8 × 8 × 4)
Initialize: k ($0 \leqslant k \leqslant$ T)
Extract discriminative feature representations from the feature map using spatio-temporal neural network
Obtain the CPM (T × 83) from the last dense layer
  If **train:**
    Calculate the CTC loss according to the CPM and target label
      $$\mathrm{CTC}Loss = -\log p\,(L \mid X)$$
    Output: CTC loss
  end if
  If inference:
    If greedy search decoding:
      Find every character x_t corresponding to the maximum probability value of each row of CPM, then the predicted sequence X could be obtained.
      Apply many-to-one function β to X and get t-he decoded sequence.
      Output: the decoded sequence $\beta(X)$ by the LM processing
    end if
    If beam search decoding:
      while $t \leqslant T$ **do**
        Find top k probability: $p\,(\pi \mid X) = \prod_1^t y_{\pi_t}$ at the t frame
        Apply many-to-one function $\beta(\pi)$ to the k predicted sequence and sum the probability with the same decoded sequence.
      end while
      Output: a total of k decoded sequence $\beta(\pi)$ after the LM processing
    end if
  end if

was applied to optimize all parameters with a full batch size [42]. The networks were trained with a 0.01 learning rate for 500 epochs. The proposed method was established on Keras framework by Python 3.6 running on a NVIDIA GeForce GPU of RTX3060.

C. Evaluation Criteria and Comparison Methods

The model was tested in a subject-specific manner. The data from each subject were divided into a training set, a validation set and a testing set, with a proportion of 60%, 20%, and 20%, respectively. In addition, a five-fold cross-validation strategy was adopted to make full use of data.

Character error rate (CER) is a classic metric to evaluate the performance of Chinese ASR. In this paper, we labeled each syllable by one Chinese character. Comparing the decoded sequence with a sequence of character labels in the actual phrase, three kinds of errors could be reported, namely insertion, deletion and substitution. These errors were counted to compute CER according to Eq. 4, to evaluate the decoding performance at the syllable/character level. A smaller CER indicates better performance.

TABLE I
CONFIGURATIONS OF THE COMPARISON METHODS

Comparison methods		Settings
Decoding method	BiL-CTC	BiLSTM (256), BN, Dropout (0.2), BiLSTM (256), BN, Dropout (0.2), BiLSTM (128), BN, Dropout (0.2).
Classification Methods	RF	700 decision trees [22]
	SVM	a radial basis function kernel [16]
	LDA	Linear discriminant analysis [21]
	LR	500 iterations
	CNN	CNN (32), BN, Dropout (0.3), Maxpooling, CNN (16), BN, Dropout (0.3), Maxpooling, Dense (1024), Dense (128), Dense (33).
	BiLSTM	BiLSTM (512), BN, Dropout (0.3), BiLSTM (256), BN, Dropout (0.3), BiLSTM (128), BN, Dropout (0.3). Dense (33)

$$CER = \frac{Insertions + Substitutions + Deletions}{Totalcharactersinphrases} \quad (4)$$

Besides, phrase classification accuracy (PCA) was designed to evaluate classification performance at the entire phrase level. It is defined as:

$$PCA = \frac{Numberofcorrectlyrecognizedsamples}{Numberofallphrasesamples} \quad (5)$$

The effectiveness of the proposed method was verified from perspectives of both decoding and classification, and some conventional methods for myoelectric pattern recognition were also selected and carried out for performance comparison. For decoding syllable/character sequences, another method was designed following the idea of ablation experiment. In this method, the spatial block for characterizing spatial information was removed, to identify how spatial features mined by the proposed method contribute to the decoding performance. Both modules remained in the neural networks including the temporal block (i.e., the BiLSTM module) and the CTC decoder, and this method is termed BiL-CTC. Specifically, the LM was not considered to directly reflect the decoding performance of the neural networks. Likewise, corresponding part of the proposed method can also be denoted as CBiL-CTC by adding the spatial block of the CNN module. We optimized neural network settings of the BiL-CTC method. Finally, three BiLSTM layers were adopted with the neuron number set to 256, 256 and 128, respectively. Each layer was followed by a BN layer and a dropout layer (a learning rate of 0.2). The original feature map (T × 8 × 8 × 4) were reshaped into the size of T × 256, suitable for being fed into the BiL-CTC method.

In addition, several classifiers commonly explored in the literature were selected as benchmark classification methods for performance comparison in terms of the PCA, including the CNN, BiLSTM, RF, SVM, LDA and logistic regression

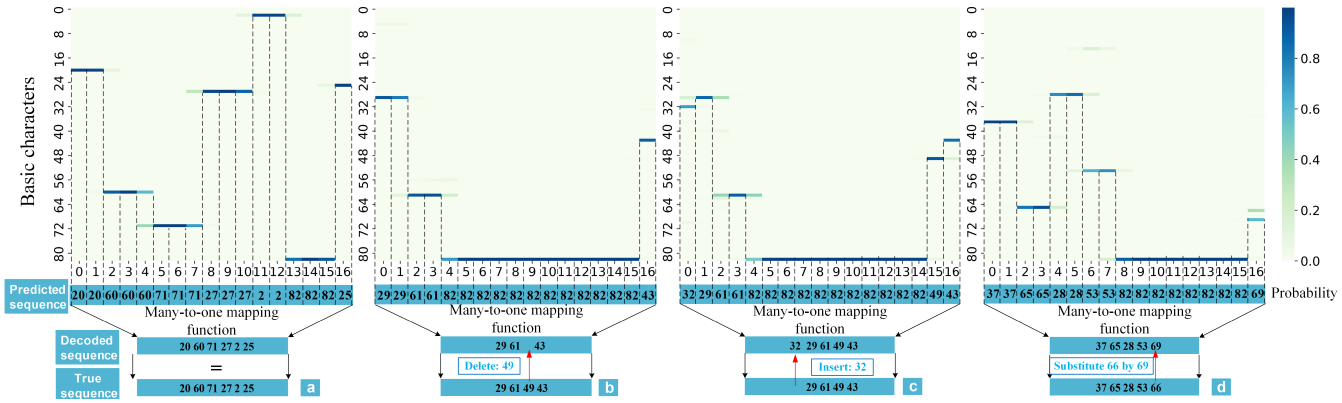

Fig. 4. The syllable-level decoding of the CBiL-CTC method using greedy search algorithm. The horizontal axis represents time frame and the vertical axis is the index of the basic Chinese syllables as listed in Fig. 1. The color bar represents probability of a syllable to be decoded by the proposed method.

(LR) methods. For the CNN method, we converted the original 64 channel signals into images in the form of 64 × T × 4, so that the convolution kernel was able to extract the spatio-temporal features simultaneously from HD-sEMG signals, where T was the number of frames and 4 represented the number of features in each frame. In the CNN network structure, two convolutional layers were applied with 32 and 16 convolution kernels, respectively. Each convolutional layer was followed by a BN layer, maxpooling layer and dropout layer with a learning rate of 0.3. We designed three bidirectional LSTM layers for the BiLSTM method and each layer is followed by a BN layer. The number of neurons in each layer was optimally set to 512, 256 and 128 respectively. The input dimension of the BiLSTM method was the same with the BiL-CTC method. We selected 700 decision trees for the RF method, and the SVM method with a linear kernel was applied in this study. For the LR method, 500 iterations were set. These methods (i.e., the RF, SVM and LR method) required data input in a one-dimensional feature vector, by flattening the features from original 64 channels into a one-dimensional vector. Both the network design and the parameter setting were verified by sufficient pretests towards optimal performance. Specific configurations are reported in Table I. Specifically, the BiL-CTC decoding method incorporated with the LM was also used for classification performance comparison, termed BiL-CTC-LM. Thus, the proposed method can be denoted as CBiL-CTC-LM.

A one-way repeated-measures ANOVA was applied to CER to examine the effect of the decoding method (2 levels: the BiL-CTC method and the CBiL-CTC method). Another one-way repeated-measures ANOVA was applied to PCA toexamine the effect of classification methods (8 levels: the RF, LR, LDA, SVM, CNN, BiLSTM, BiL-CTC-LM and proposed CBiL-CTC-LM methods). Before any ANOVA was applied, the normality tests were applied using the Shapiro-Wilk test. If possible, post hoc multiple pairwise comparisons were conducted with Bonferroni corrections. All of the statistical analyses were performed using SPSS software (22.0 version, SPSS Inc.) and the significance level was set to 0.05.

III. Results

A. Visualization of Syllable-Level Speech Decoding

To shed light on how the proposed method decoded the articulatory features at each frame, Fig. 4 presents four cases of decoding syllable-level speech with the greedy search algorithm using the CBiL-CTC method from a representative subject, where the frame length and increment were selected as 200ms and 120ms, respectively. In Fig. 4, the predicted sequence showed the intuitive decoding results at every frame, and then a many-to-one function was adopted to obtain the decoded sequence by removing duplicated syllables and *blank* symbols. Fig. 4(a) illustrates a correctly decoded sequence (T33). Fig. 4(b-d) shows three types of syllable-level errors in the decoded sequences, namely deletion, insertion and substitution with respect to the true sequences (T22, T22, T24), respectively. In addition, more errors even in different types might take place simultaneously in the decoded sequence with respect to one true sequence.

B. End-to-End Decoding Performance

Table II compares the decoding performance in terms of CER between the CBiL-CTC method and the BiL-CTC method using the beam search algorithm under different frame lengths and increments. We can notice that the CBiL-CTC method achieves the minimum CER value (3.11 ± 1.46%) when frame length and increment are 200ms and 180ms respectively, showing superior performance than the BiL-CTC method (4.76 ± 1.94%) when both frame length and increment are set to 180ms, with statistical significance ($p = 0.026 < 0.05$). The Shapiro-Wilk test prior to the ANOVA confirmed that both groups of data followed normal distributions ($p > 0.05$).

C. Effect of The Language Model for Phrase Classification

Fig. 5 reports the mean PCAs using both the BiL-CTC and CBiL-CTC methods, and the combination of each method with the LM (i.e., the BiL-CTC-LM method and CBiL-CTC-LM method). It is evident that the use of LM leads to a PCA

TABLE II

CER (%) OF THE CBiL-CTC AND BiL-CTC METHODS CALCULATED AS A FUNCTION OF THE FRAME LENGTH AND INCREMENT. THE CER IS AVERAGED ACROSS ALL SUBJECTS WITH THE STANDARD DEVIATION

Frame increment	Method	Frame length						
		120	140	160	180	200	220	240
100	CBiL-CTC	4.74 ± 2.13	4.39 ± 1.61	5.04 ± 1.94	5.85 ± 3.29	4.41 ± 2.91	4.41 ± 2.34	5.20 ± 1.66
	BiL-CTC	7.42 ± 5.52	7.62 ± 3.66	8.54 ± 4.56	8.26 ± 6.82	8.88 ± 4.53	8.27 ± 5.09	8.93 ± 4.74
120	CBiL-CTC	4.36 ± 2.46	4.09 ± 2.07	4.11 ± 2.15	4.22 ± 2.23	4.81 ± 1.85	4.32 ± 2.66	4.39 ± 1.27
	BiL-CTC	5.16 ± 2.33	5.49 ± 2.70	6.61 ± 2.99	7.70 ± 2.67	9.55 ± 4.81	7.58 ± 1.40	7.64 ± 3.24
140	CBiL-CTC	--	3.72 ± 2.38	3.91 ± 2.25	3.36 ± 2.06	3.82 ± 1.52	3.92 ± 1.95	3.36 ± 1.83
	BiL-CTC	--	5.45 ± 3.37	6.66 ± 1.57	6.65 ± 3.29	6.27 ± 1.83	9.06 ± 3.78	7.71 ± 3.03
160	CBiL-CTC	--	--	4.13 ± 2.91	3.81 ± 2.48	3.72 ± 2.03	3.35 ± 1.63	3.62 ± 1.75
	BiL-CTC	--	--	6.14 ± 2.6	6.51 ± 3.90	7.49 ± 3.69	7.45 ± 3.58	5.74 ± 1.37
180	CBiL-CTC	--	--	--	3.41 ± 1.50	**3.11 ± 1.46**	4.31 ± 2.23	3.86 ± 2.24
	BiL-CTC	--	--	--	**4.76 ± 1.94**	6.30 ± 2.95	6.17 ± 1.87	6.09 ± 2.08
200	CBiL-CTC	--	--	--	--	3.45 ± 2.24	3.43 ± 2.32	4.36 ± 2.85
	BiL-CTC	--	--	--	--	6.56 ± 3.63	5.53 ± 2.18	5.58 ± 2.31

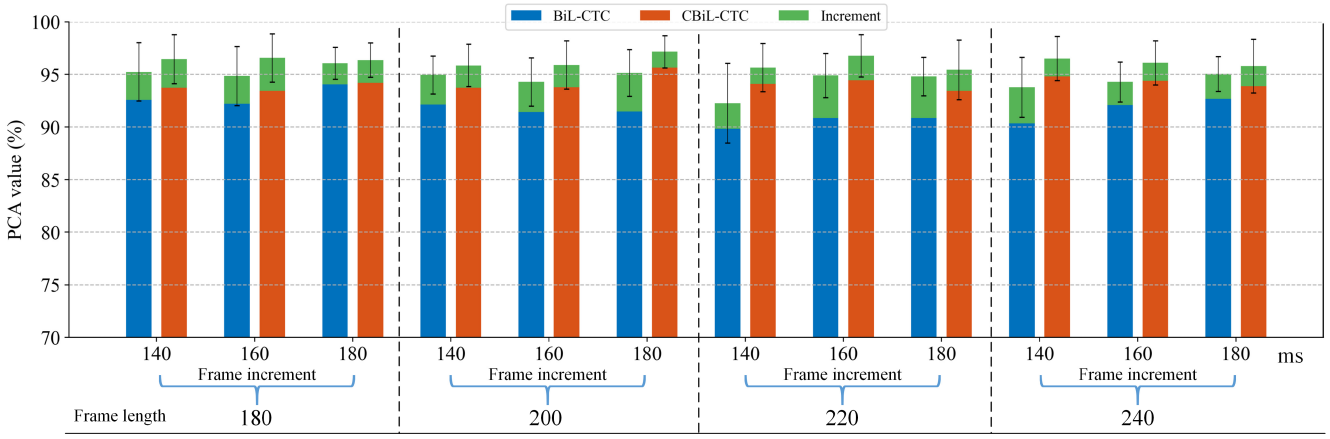

Fig. 5. The PCAs averaged across all subjects under different frame lengths and increments, using the BiL-CTC method (blue bar) and the CBiL-CTC method (red bar), and each method incorporated with the LM. The green bar indicates PCA improvement derived from the LM under each condition. Error bars represent standard deviations.

improvement, which is highlighted in green bars under different conditions. The BiL-CTC-LM method had the optimal PCA of 96.06 ± 1.52% at both the frame length and increment set to 180ms, including a PCA lift of 2.02% by the LM. At a frame length of 200 ms and a frame increment of 180ms, the proposed CBiL-CTC-LM method yielded the highest PCA of 97.17 ± 1.53%, including a PCA improvement of 1.51%.

D. Phrase Classification Performance

Table III lists the phrase classification performance in terms of PCA using the proposed method and other classification methods, under different conditions of the frame length and increment, respectively. Four methods with deep neural networks (i.e., the CNN, BiLSTM, BiL-CTC-LM and proposed CBiL-CTC-LM method) exhibited relatively higher PCAs reaching to 91.72 ± 4.63%, 93.23 ± 2.99%, 96.06 ± 1.52% and 97.17 ± 1.53%, respectively. Four other non-deep classification methods had PCAs lower than 90% (i.e. 87.98 ± 5.26% for the SVM method, 84.54 ± 2.17% for the RF method, 88.38 ± 4.49% for the LDA method and 85.76 ± 5.13% for the LR method).

Fig. 6 shows the boxplots of the PCAs for all subjects using eight different methods, when the frame settings were customized corresponding to each method towards optimal performance. After the Shapiro-Wilk test reported a normal distribution for any data group ($p > 0.05$), the ANOVA revealed that the proposed method significantly outperformed other methods ($p < 0.05$) except the BiL-CTC-LM method ($p = 0.84$).

IV. DISCUSSIONS

Conventional SSR studies focus on the classification of a finite number of words or phrases. When a phrase or a sequence of words was regarded as a single pattern, the fine-grained phoneme or syllable relevance might be ignored. Such fine-grained information needs to be well characterized, thus leading to improved SSR performance. This paper presents an efficient method for decoding continuous syllable

TABLE III

PCAs (%) AVERAGED OVER ALL SUBJECTS USING THE PROPOSED METHOD AND OTHER CLASSIFICATION METHODS, UNDER DIFFERENT CONDITIONS OF THE FRAME LENGTH AND INCREMENT

Frame increment	Method	Frame length							
		180	p	200	p	220	p	240	p
140	CNN	91.41 ± 4.48	*	91.02 ± 4.62	0.228	90.96 ± 4.04	**	91.72 ± 4.63	**
	BiLSTM	93.03 ± 3.52	0.057	93.18 ± 3.82	0.575	93.23 ± 2.99	0.364	92.88 ± 3.49	*
	SVM	85.86 ± 6.10	***	85.86 ± 5.99	***	86.06 ± 5.63	***	87.47 ± 5.14	***
	RF	82.48 ± 6.07	***	82.63 ± 4.94	***	83.30 ± 5.34	***	84.54 ± 2.17	***
	LDA	88.03 ± 4.82	***	88.33 ± 4.47	***	88.23 ± 4.96	**	88.23 ± 5.16	***
	LR	84.39 ± 6.77	***	85.76 ± 5.13	***	83.54 ± 6.47	***	84.39 ± 6.00	***
	BiL-CTC-LM	95.25 ± 2.77	1.000	94.95 ± 1.79	0.770	92.27 ± 3.79	0.059	93.79 ± 2.85	0.228
	CBiL-CTC-LM	**96.46 ± 2.34**	-	**95.86 ± 2.01**	-	**95.66 ± 2.29**	-	**96.52 ± 2.10**	-
160	CNN	90.40 ± 6.05	*	90.51 ± 5.14	*	91.11 ± 4.07	***	91.11 ± 4.09	**
	BiLSTM	93.03 ± 3.30	*	92.93 ± 3.40	0.069	92.42 ± 4.07	**	92.88 ± 2.89	*
	SVM	87.02 ± 5.35	***	87.68 ± 4.98	**	87.78 ± 4.88	***	87.98 ± 5.26	**
	RF	81.84 ± 5.61	***	82.95 ± 5.88	***	83.57 ± 5.18	***	83.94 ± 5.09	***
	LDA	88.38 ± 4.49	***	88.08 ± 4.86	**	87.17 ± 5.37	***	87.88 ± 4.81	***
	LR	84.55 ± 5.88	***	84.75 ± 5.45	***	85.05 ± 6.43	***	85.35 ± 6.35	***
	BiL-CTC-LM	94.85 ± 2.81	0.429	94.29 ± 2.30	0.364	94.90 ± 2.10	0.784	94.29 ± 1.89	0.131
	CBiL-CTC-LM	**96.57 ± 2.31**	-	**95.91 ± 2.29**	-	**96.77 ± 2.02**	-	**96.11 ± 2.10**	-
180	CNN	90.66 ± 4.70	**	90.45 ± 5.73	**	90.56 ± 4.79	*	90.30 ± 5.07	*
	BiLSTM	92.37 ± 4.28	*	91.97 ± 4.05	**	92.22 ± 3.88	0.105	92.17 ± 4.34	0.074
	SVM	86.87 ± 4.94	***	86.97 ± 5.46	***	86.62 ± 5.08	***	85.56 ± 5.93	***
	RF	81.50 ± 5.82	***	81.78 ± 5.09	***	82.48 ± 5.25	***	83.33 ± 5.26	***
	LDA	87.88 ± 5.31	**	87.58 ± 5.32	***	87.22 ± 5.17	**	87.88 ± 4.79	**
	LR	85.56 ± 5.96	***	84.95 ± 5.50	***	84.70 ± 6.22	***	83.84 ± 6.31	***
	BiL-CTC-LM	96.06 ± 1.52	0.997	95.15 ± 2.21	0.055	94.80 ± 1.83	0.979	95.05 ± 1.66	0.932
	CBiL-CTC-LM	**96.36 ± 1.64**	-	**97.17 ± 1.53**	-	**95.45 ± 2.83**	-	**95.81 ± 2.56**	-

Note: Statistical significant is reported for pairwise comparison between the proposed (CBiL-CTC-LM) method and any other method. Significance levels $p < 0.05$, $p < 0.01$ and $p < 0.001$ are denoted by *, ** and ***, respectively.

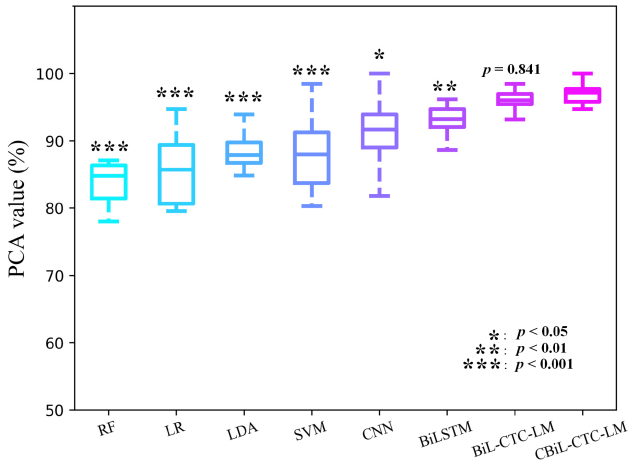

Fig. 6. The boxplot of PCAs for all subjects using eight different methods, when the frame settings were customized for each method towards the highest PCA. The statistical significance was reported for comparison between the proposed method and each of other seven methods.

sequence in silent speech from HD-sEMG recordings, using spatio- temporal end-to-end neural networks.

From an intuitive view of the syllable-level decoding process using the CBiL-CTC part of the proposed method in Fig. 4, we can notice that each frame had a syllable/character decision (according to the maximum of the probabilities of the frame data belonging to all possible syllables) in the interference stage. In the original sequence of decoded syllables, a single syllable can be repeated consecutively over multiple frames. This can be explained by the fact that each frame may not precisely cover one syllable to be articulated. It is possible that a syllable articulation may last over multiple frames. Such a finding indicates the necessity of applying a many-to-one function on the originally decoded sequence to merge repetitions of consecutive non-blank syllables and to remove the blank symbols, as described in our method. Besides, some occasional syllable errors (*i.e.*, deletions, insertions and substitutions as shown in Fig. 4(b-d) found in the decoded sequence further confirms usability of the LM in rectifying these errors, which is consistent with previous studies in natural language processing [43], [44], [45]. The performance improvement by the simple LM conducted in our study was demonstrated as accuracy increments for phrase classification (see Fig. 5).

When considering the syllable-level decoding performance, both the BiL-CTC method and the CBiL-CTC method exhibited good CERs lower than 5% (Table II), indicating the effectiveness of understanding contextual semantic information specifically by both the BiLSTM and CTC modules. This finding is consistent with most of previous studies in the fields of speech recognition and handwriting recognition [46], [47], [48]. By comparing both decoding methods, the superior performance in terms of significantly lower CER ($3.11 \pm 1.46\%$, $p < 0.05$) directly demonstrated positive effect of the CNN module for enhanced spatial feature characterization included in the proposed method. Consistent findings have been reported in advanced myoelectric control studies using the HD-sEMG recordings [23], [49], [50]. This study further

confirms usability of characterizing important spatial information from the HD-sEMG data in decoding silent speech.

When considering the phrase classification performance, conventional classifiers (*i.e.*, the CNN, BiLSTM, SVM, RF, LDA and LR methods) showed acceptable accuracies over 80%, as shown in Table III and Fig. 6. These findings were consistent with many previous reports [16], [20], [21], [22]. On this basis, the proposed CBiL-CTC-LM method working in a syllable-level decoding way evidently improved the accuracy to 95%, with statistical significance ($p < 0.05$). Such promising improvement can be attributed into powerful capability of the CTC decoder in well characterizing semantic relevance between syllables in sequential silent speech data. By contrast, conventional classification methods failed to show good ability in describing syllable sequences, being unsuitable for continuous SSR systems. Furthermore, in the CTC algorithm, a initial sequence of syllable decisions made on every input data frame was then refined by a many-to-one mapping function to meet the actual syllable sequence, as shown in Fig. 4. This end-to-end decoding way allows syllable labels not to be strictly aligned to the data stream in the training dataset, thus overcoming the well-known time-alignment challenge and facilitating practical SSR.

It is worth noting that both the BiL-CTC-LM method and the CBiL-CTC-LM method achieved comparable PCAs over 96% (Fig. 6, $p = 0.841$), although the BiL-CTC method had inferior decoding performance with a significantly higher CER than the CBiL-CTC method ($p = 0.026$). This finding indicates powerful capability of the LM that compensate the performance gap between both the BiL-CTC and CBiL-CTC decoding methods. Please also note that the BiL-CTC-LM method really had many advantageous aspects that are the same as those of the proposed method, including the CTC decoder as discussed above. Another reason for explaining why the use of spatial block in the proposed method failed to improve PCA was the limited size of corpus used in the current study. Advanced characterization of HD-sEMG spatial information is expected to benefit precise syllable identification, thus resulting in improved decoding of silent speech given a large size of corpus.

Besides, other limitations of the current study are summarized. A conditional independence is assumed in the CTC algorithm, that is, the frames are independent of each other. This assumption is unfavorable for continuous SSR, because contextual semantics during speech can be reflected by relevance among consecutive data frames, carrying abundant temporal information. The ongoing work is to find some methods (e.g., attention mechanism [51], [52]) to improve the output of the CTC algorithm. These abovementioned topics are the directions of our future work.

V. Conclusion

This study presents an end-to-end method to achieve sEMG-based SSR. In the proposed method, the CNN-BiLSTM neural network was applied to extract discriminative spatio-temporal feature representations and the CTC decoder was adopted for syllable-level decoding. Subsequently, the final decoded results were further improved by the LM. The proposed method outperformed other benchmark methods including both syllable-level decoding methods and phrase classification methods, demonstrating a promising capacity to decode sEMG-based SSR.

References

[1] M. Ganzeboom, E. Yılmaz, C. Cucchiarini, and H. Strik, "On the development of an ASR-based multimedia game for speech therapy: Preliminary results," in *Proc. ACM Workshop Multimedia Pers. Health Health Care*, Oct. 2016, pp. 3–8, doi: 10.1145/2985766.2985771.

[2] M. Katore and M. R. Bachute, "Speech based human machine interaction system for home automation," in *Proc. IEEE Bombay Sect. Symp. (IBSS)*, Sep. 2015, pp. 1–6, doi: 10.1109/IBSS.2015.7456634.

[3] G. Muhammad, "Automatic speech recognition using interlaced derivative pattern for cloud based healthcare system," *Cluster Comput.*, vol. 18, no. 2, pp. 795–802, Jun. 2015, doi: 10.1007/s10586-015-0439-7.

[4] K. Wu, C. Zhang, X. Wu, D. Wu, and X. Niu, "Research on acoustic feature extraction of crying for early screening of children with autism," in *Proc. 34rd Youth Academic Annu. Conf. Chin. Assoc. Autom. (YAC)*, Jun. 2019, pp. 290–295, doi: 10.1109/YAC.2019.8787725.

[5] M. Kim, B. Cao, T. Mau, and J. Wang, "Speaker-independent silent speech recognition from flesh-point articulatory movements using an LSTM neural network," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2323–2336, Dec. 2017.

[6] N. Kimura, M. Kono, and J. Rekimoto, "SottoVoce: An ultrasound imaging-based silent speech interaction using deep neural networks," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2019, pp. 1–11, doi: 10.1145/3290605.3300376.

[7] A. D. Scott, M. Wylezinska, M. J. Birch, and M. E. Miquel, "Speech MRI: Morphology and function," *Phys. Medica*, vol. 30, no. 6, pp. 604–618, Sep. 2014, doi: 10.1016/j.ejmp.2014.05.001.

[8] R. Bowden et al., "Recent developments in automated lip-reading," *Proc. SPIE*, vol. 8901, Oct. 2013, Art. no. 89010J, doi: 10.1117/12.2029464.

[9] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline, "Development of sEMG sensors and algorithms for silent speech recognition," *J. Neural Eng.*, vol. 15, no. 4, Aug. 2018, Art. no. 046031, doi: 10.1088/1741-2552/aac965.

[10] B. J. Betts, C. Jorgensen, and M. Field, "Small vocabulary recognition using surface electromyography in an acoustically harsh environment," *J. Hum.-Comput. Interact.*, vol. 18, no. 6, pp. 1242–1259, 2006.

[11] 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, Language Process.*, vol. 25, no. 12, pp. 2386–2398, Dec. 2017, doi: 10.1109/TASLP.2017.2740000.

[12] M. Wand and J. Schmidhuber, "Deep neural network frontend for continuous EMG-based speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, vols. 8–12, 2016, pp. 3032–3036, doi: 10.21437/Interspeech.2016-340.

[13] A. Porbadnigk, M. Wester, J. Calliess, T. Schultz, and A. Agcl, "EEG-based speech recognition—Impact of temporal effects," in *Proc. Int. Conf. Bio-Inspired Syst. Signal Process.*, 2011, pp. 376–381, doi: 10.5220/0001554303760381.

[14] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, Apr. 2019, doi: 10.1038/s41586-019-1119-1.

[15] D. Gaddy and D. Klein, "Digital voicing of silent speech," 2020, *arXiv:2010.02960*.

[16] J. He, X. Wang, X. Zhang, B. Wang, Q. Li, and C. Qiu, "Unvoiced speech recognition algorithm based on myoelectric signal," in *Proc. 12th Int. Conf. Mach. Learn. Comput.*, Feb. 2020, pp. 450–456, doi: 10.1145/3383972.3384029.

[17] H. S. Lo and S. Q. Xie, "Exoskeleton robots for upper-limb rehabilitation: State of the art and future prospects," *Med. Eng. Phys.*, vol. 34, no. 3, pp. 261–268, Apr. 2012, doi: 10.1016/j.medengphy.2011.10.004.

[18] P. Parker, K. Englehart, and B. Hudgins, "Myoelectric signal processing for control of powered limb prostheses," *J. Electromyogr. Kinesiol.*, vol. 16, no. 6, pp. 541–548, Dec. 2006, doi: 10.1016/j.jelekin.2006.08.006.

[19] R. A. R. C. Gopura, D. S. V. Bandara, K. Kiguchi, and G. K. I. Mann, "Developments in hardware systems of active upper-limb exoskeleton robots: A review," *Robot. Auto. Syst.*, vol. 75, pp. 203–220, Jan. 2016, doi: 10.1016/j.robot.2015.10.001.

[20] X. Wang et al., "A pilot study on the performance of time-domain features in speech recognition based on high-density sEMG," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 19–22, doi: 10.1109/EMBC46164.2021.9630541.

[21] M. Zhu et al., "The effects of electrode locations on silent speech recognition using high-density sEMG," in *Proc. IEEE Int. Workshop Metrol. Ind. 4.0 IoT*, Jun. 2020, pp. 345–348, doi: 10.1109/MetroInd4.0IoT48571.2020.9138289.

[22] M. Zhang, W. Zhang, B. Zhang, Y. Wang, and G. Li, "Feature selection of mime speech recognition using surface electromyography data," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2019, pp. 3173–3178, doi: 10.1109/CAC48633.2019.8996646.

[23] S. Zhang, X. Zhang, S. Cao, X. Gao, X. Chen, and P. Zhou, "Myoelectric pattern recognition based on muscle synergies for simultaneous control of dexterous finger movements," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 4, pp. 576–582, Aug. 2017, doi: 10.1109/THMS.2017.2700444.

[24] X. Zhang and P. Zhou, "High-density myoelectric pattern recognition toward improved stroke rehabilitation," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 6, pp. 1649–1657, Jun. 2012, doi: 10.1109/TBME.2012.2191551.

[25] X. Zhang, L. Wu, B. Yu, X. Chen, and X. Chen, "Adaptive calibration of electrode array shifts enables robust myoelectric control," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 7, pp. 1947–1957, Jul. 2020, doi: 10.1109/TBME.2019.2952890.

[26] M. Wand, C. Schulte, M. Janke, and T. Schultz, "Array-based electromyographic silent speech interface," in *Proc. Int. Conf. Bio-Inspired Syst. Signal Process. (BIOSIGNALS)*, 2013, pp. 89–96, doi: 10.5220/0004252400890096.

[27] M. Janke and L. Diener, "EMG-to-speech: Direct generation of speech from facial electromyographic signals," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2375–2385, Dec. 2017, doi: 10.1109/TASLP.2017.2738568.

[28] M. Zhu et al., "Towards optimizing electrode configurations for silent speech recognition based on high-density surface electromyography," *J. Neural Eng.*, vol. 18, no. 1, Feb. 2021, Art. no. 016005, doi: 10.1088/1741-2552/abca14.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[30] S. C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards continuous speech recognition using surface electromyography," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, vols. 1–5, 2006, pp. 573–576, doi: 10.1021/ol800080w.

[31] J.-Y. Hsu, Y.-J. Chen, and H.-Y. Lee, "Meta learning for end-to-end low-resource speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7844–7848, doi: 10.1109/ICASSP40776.2020.9053112.

[32] J. Luo, J. Wang, N. Cheng, and J. Xiao, "Loss prediction: End-to-end active learning approach for speech recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–7, doi: 10.1109/IJCNN52387.2021.9533839.

[33] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification," in *Proc. 23rd Int. Conf. Mach. Learn.*, vol. 148, 2006, pp. 369–376, doi: 10.1145/1143844.1143891.

[34] H. Ye et al., "Attention bidirectional LSTM networks based mime speech recognition using sEMG data," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2020, pp. 3162–3167, doi: 10.1109/SMC42975.2020.9282863.

[35] Y. Zhou, R. Ouyang, and X. Wu, "MI-EEG temporal information learning based on one-dimensional convolutional neural network," in *Proc. 35th Youth Academic Annu. Conf. Chin. Assoc. Autom. (YAC)*, Oct. 2020, pp. 499–504, doi: 10.1109/YAC51587.2020.9337703.

[36] D. Lee et al., "Long short-term memory recurrent neural network-based acoustic model using connectionist temporal classification on a large-scale training corpus," *China Commun.*, vol. 14, no. 9, pp. 23–31, Sep. 2017, doi: 10.1109/CC.2017.8068761.

[37] Y. Wang et al., "Silent speech decoding using spectrogram features based on neuromuscular activities," *Brain Sci.*, vol. 10, no. 7, pp. 1–14, 2020, doi: 10.3390/brainsci10070442.

[38] A. Al-Timemy, R. Khushaba, G. Bugmann, and J. Escudero, "Improving the performance against force variation of EMG controlled multifunctional upper-limb prostheses for transradial amputees," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 6, pp. 650–661, Jun. 2016, doi: 10.1109/TNSRE.2015.2445634.

[39] B. Hudgins, P. Parker, and R. N. Scott, "A new strategy for multifunction myoelectric control," *IEEE Trans. Biomed. Eng.*, vol. 40, no. 1, pp. 82–94, Jan. 1993, doi: 10.1109/10.204774.

[40] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 448–456.

[41] E. S. Ristad and P. N. Yianilos, "Learning string-edit distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 5, pp. 522–532, May 1998, doi: 10.1109/34.682181.

[42] T. Dozat, "Incorporating Nesterov momentum into ADAM," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–4.

[43] D. Linares, J.-M. Benedí, and J.-A. Sánchez, "A hybrid language model based on a combination of N-grams and stochastic context-free grammars," *ACM Trans. Asian Lang. Inf. Process.*, vol. 3, no. 2, pp. 113–127, Jun. 2004, doi: 10.1145/1034780.1034783.

[44] D. C. Cavalieri, S. E. Palazuelos-Cagigas, T. F. Bastos-Filho, and M. Sarcinelli-Filho, "Combination of language models for word prediction: An exponential approach," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1481–1494, Sep. 2016, doi: 10.1109/TASLP.2016.2547743.

[45] T. Brychcín and M. Konopík, "Semantic spaces for improving language modeling," *Comput. Speech Lang.*, vol. 28, no. 1, pp. 192–209, Jan. 2014, doi: 10.1016/j.csl.2013.05.001.

[46] S. Adya, V. Garg, S. Sigtia, P. Simha, and C. Dhir, "Hybrid transformer/CTC networks for hardware efficient voice triggering," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Oct. 2020, pp. 1–5, doi: 10.21437/Interspeech.2020-1330.

[47] M. A. M. Sakib, O. Sharif, and M. M. Hoque, "Offline Bengali handwritten sentence recognition using BiLSTM and CTC networks," in *Proc. Int. Conf. Internet Things Connected Technol.*, 2021, pp. 158–168, doi: 10.1007/978-3-030-76736-5_15.

[48] D. Raval, V. Pathak, M. Patel, and B. Bhatt, "End-to-end automatic speech recognition for Gujarati," in *Proc. 17th Int. Conf. Natural Lang. Process. (ICON)*, 2020, pp. 409–419.

[49] J. He and X. Zhu, "Combining improved gray-level co-occurrence matrix with high density grid for myoelectric control robustness to electrode shift," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 9, pp. 1539–1548, Sep. 2017, doi: 10.1109/TNSRE.2016.2644264.

[50] J. Cheng, F. Wei, C. Li, Y. Liu, A. Liu, and X. Chen, "Position-independent gesture recognition using sEMG signals via canonical correlation analysis," *Comput. Biol. Med.*, vol. 103, pp. 44–54, Dec. 2018, doi: 10.1016/j.compbiomed.2018.08.020.

[51] J. Salazar, K. Kirchhoff, and Z. Huang, "Self-attention networks for connectionist temporal classification in speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 7115–7119, doi: 10.1109/ICASSP.2019.8682539.

[52] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2015, pp. 1–9.