Article

Sequence-to-Sequence Voice Reconstruction for Silent Speech in a Tonal Language

Huiyan Li [1], Haohong Lin [1], You Wang [1,*], Hengyang Wang [1], Ming Zhang [1], Han Gao [1], Qing Ai [1], Zhiyuan Luo [2] and Guang Li [1]

[1] State Key Laboratory of Industrial Control Technology, Institute of Cyber Systems and Control, Zhejiang University, Hangzhou 310027, China; huiyanli@zju.edu.cn (H.L.); lhh2017@zju.edu.cn (H.L.); 11432014@zju.edu.cn (H.W.); drystan@zju.edu.cn (M.Z.); gao_han@zju.edu.cn (H.G.); aiqing@zju.edu.cn (Q.A.); guangli@zju.edu.cn (G.L.)

[2] Department of Computer Science, Royal Holloway, University of London, Egham Hill, Egham TW20 0EX, Surrey, UK; zhiyuan.luo@rhul.ac.uk

* Correspondence: king_wy@zju.edu.cn

Abstract: Silent speech decoding (SSD), based on articulatory neuromuscular activities, has become a prevalent task of brain–computer interfaces (BCIs) in recent years. Many works have been devoted to decoding surface electromyography (sEMG) from articulatory neuromuscular activities. However, restoring silent speech in tonal languages such as Mandarin Chinese is still difficult. This paper proposes an optimized sequence-to-sequence (Seq2Seq) approach to synthesize voice from the sEMG-based silent speech. We extract duration information to regulate the sEMG-based silent speech using the audio length. Then, we provide a deep-learning model with an encoder–decoder structure and a state-of-the-art vocoder to generate the audio waveform. Experiments based on six Mandarin Chinese speakers demonstrate that the proposed model can successfully decode silent speech in Mandarin Chinese and achieve a character error rate (CER) of 6.41% on average with human evaluation.

Keywords: silent speech; electromyography (EMG); neuromuscular signal; sequence-to-sequence (Seq2Seq)

1. Introduction

Silent speech decoding (SSD) is one of the most popular areas of brain–computer interface (BCI) research, which makes it possible for humans to interact with their surroundings and express their inner minds without speaking words [1,2]. SSD aims at detecting biological speech-related activities (instead of acoustic data) and decoding the thoughts of humans using physiological measurements.

Speech-related signals detected by physiological measurements are defined as biosignals [3]. The typical physiological measurements are obtained by using sensors to capture biosignals from the brain [4], e.g., electrocorticography (ECoG) [5–7] and electroencephalography (EEG) [8,9]. However, these devices for biosignal acquisition have several disadvantages. ECoG is invasive and probably has surgical complications [10]; EEG has no harmful side effects, but the signal processing of EEG is difficult for practical use [2]. Acquisition of neuromuscular signals is a promising way to decode speech-related activity [3].

Surface electromyography (sEMG), which is non-invasive and convenient to apply in practical applications, can be used to acquire the control signals that are transferred from the cortex to the facial muscles and then decode the silent speech [11]. In addition, the neural pathways from the brain to muscle can act as primary filters and encoders [12], and EMG has lower channel requirements [2]. Electromagnetic articulography (EMA) sensors [13] and optical imaging of the tongue and lips [14] are also often used in SSD to record invisible speech articulators. However, they could not work in the absence of articulator movement.

Existing studies on SSD can be divided into two categories: *biosignal-to-text* and *biosignal-to-voice* [3]. The former can be regarded as a kind of classification task, while the latter is a kind of regression task [6]. Considering that the biosignal-to-text approaches may lose some information about the speaker's personality and emotion during processing, the two-step approach "biosignal-to-text-to-voice" is too time-consuming for real-time scenarios [7]. Many works have tried to decode silent speech by reconstructing voices [3,7]. In this paper, we denote this task as sEMG-to-voice (*sEMG2V*). This technology has many applications. It will not be interfered with by external noise, which makes it remain effective in noisy environments such as factories. This technology can also help patients who are no longer able to speak due to surgical removal of their larynx due to trauma or diseases [1,15–17]. In addition, this mode is more concealed and cannot be observed through lip language analysis and other means, which offers more privacy protection.

The existing methods of *sEMG2V* in a tonal language have the following problems:

Research on SSD is mainly concentrated in non-tonal languages such as English [15,16] while SSD approaches for tonal languages are limited to solving the tasks of classification [2,18,19]. Different from non-tonal languages, the pitches, called tones in tonal languages, carry more lexical or grammatical information to distinguish one word from another [20–23]. It has been shown that tones carry no less information than vowels in Mandarin Chinese [24]. The distinctive tonal patterns of language are called tonemes [25,26] to distinguish them from phonemes. There are five tones in Mandarin Chinese, which are transcribed by letters with diacritics over vowels [27]: high level tone (first tone) such as /bā/ (eight), rising tone (second tone) such as /bá/ (to pull), dipping tone (third tone) such as /bǎ/ (to hold), high-falling tone (fourth tone) such as /bà/ (father), and neutral tone (fifth tone) such as /ba/ (an interrogative particle). The total number of tonemes, including toned vowels and consonants, is 139 in Mandarin Chinese [28], while the total number of phonemes in English, which is a non-tonal language, is 44 [29]. With the same number of datasets, Mandarin contains a larger dimension of information than English and is more difficult to decode. One study recognizes 10 Chinese words in silent speech with an accuracy of 90% [2]. To our knowledge, there is little work studying sEMG2V in tonal languages.

In addition, the sEMG-based silent speech has no time-aligned parallel audio. To provide time-aligned parallel information, dynamic time warping (DTW) [30] can be applied to obtain alignments between silent and vocal speech [15]. Recently, Gaddy and Klein [16] utilized predicted audio for DTW to achieve alignments, extract audio features in parallel, and obtain a word error rate (WER) of 68.0% in the sEMG-based silent speech. The accuracy can still be improved by finding a better approach for providing corresponding audios.

In order to address these limitations of existing *sEMG2V* methods in a tonal language, this paper proposes a novel approach based on a sequence-to-sequence (Seq2Seq) model, inspired by the tremendous success of the Seq2Seq model in text-to-speech (TTS) and voice conversion (VC) [31–34]. This technology can help solve the unparallel information between silent speech and audio. We utilize a length regulator [32] for the *sEMG2V* to obtain audio signals. The key contributions of this paper are summarized as follows:

1. The paper proposes a Seq2Seq model, the first attempt to introduce a Seq2Seq model into the *sEMG2V* task. The model extracts duration information from the alignment between sEMG-based silent speech and vocal speech. The lengths of input sequences are adjusted to match the size of output sequences. Thus, our model can generate audios from neuromuscular activities using the Seq2Seq model.
2. The model in the paper generates audios from sEMG-based silent speech by considering both vocal sEMG reconstruction loss and toneme classification loss, and uses a state-of-the-art vocoder to achieve better quality and higher accuracy of the reconstructed audios.
3. We collect an sEMG-based silent speech dataset with Mandarin Chinese and conduct extensive experiments to demonstrate that the proposed model can decode neuromuscular signals in silent speech successfully in the tonal language.

2. Data Acquisition

2.1. Recording Information

The signal from facial skin is collected by a multi-channel sEMG data recording system using standard wet surface Ag/AgCl electrodes, as described in [2]. Meanwhile, we use a headset microphone to record audio. The views of the electrodes around the face are shown in Figure 1, and the electrode positions are shown in Table 1. The electrode positions are highly correlated with vocalizing muscles and have different meanings in speech production [19]. In our case, channel 1 is differential electrodes, and the others are single electrodes. The differential electrodes can improve the common-mode rejection ratio and improve the quality of signal [18].

Table 1. Electrode Location Details.

Electrode Index	Position
1	1 cm right from the nose
2	1 cm right from corners of the mouth
3	1 cm left from the nose
4	left corner of chin
5	4 cm behind the chin

(a)   (b)   (c)
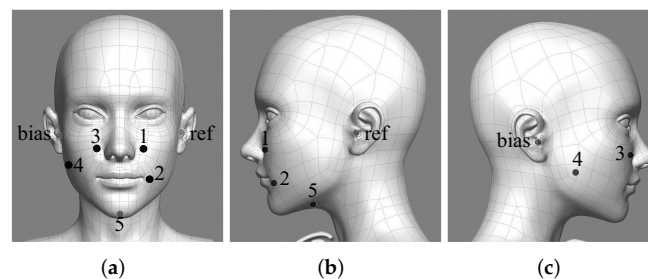
Figure 1. Three views of electrode distribution around the face and neck. (**a**) Main view, (**b**) right view, (**c**) left view.

2.2. Dataset Information

We collect the data from six native Mandarin-speaking healthy young Chinese adults with normal vision and oral expression skills. The average age of the six participants is 25. The participants are asked to clean their face before the experiment and sit still wearing electrodes and a microphone. They are trained to press the start button, read the sentences shown on the computer screen in vocal and silent mode, and press the end button. In silent mode, the participants are trained to imagine speaking sentences displayed on the computer screen as [2] shows and slight muscle motion is allowed. The dataset includes the pair of simultaneously recorded vocal sEMG ($sEMG_v$) and audio signal ($Audio_v$), and silent sEMG data (namely, $sEMG_s$). The vocal mode is recorded once, while the silent mode is repeated five times. Each recording uses phonetically balanced utterances from a Chinese corpus called AISHELL3 [35]. There are a total of 2260 words and 1373 characters in this dataset. The dataset includes six speakers, and each of them has at least 0.73 h of silent speech data, leading to 5.79 h in total. The dataset of each speaker is split into a training, validation, and testing set, with a ratio of 8:1:1 according to the number of silent utterances from each speaker, ensuring that they are phonetically balanced. Table 2 gives some statistics of each speaker. In the following, the collected dataset is denoted as $sEMG\_Mandarin$.

Table 2. Statistics of the *sEMG_Mandarin* Dataset.

Speaker Id	Sex	Silent Speech Time (Min)			Number of Utterances		
		Train	Val	Test	Train	Val	Test
1	f	52.59	6.68	6.49	680	85	85
2	m	52.65	6.70	6.26	516	64	64
3	f	56.60	7.11	6.99	716	89	89
4	m	40.77	5.11	5.01	800	100	100
5	m	40.49	5.06	5.09	600	75	75
6	m	34.85	4.29	4.40	600	75	75
Total		277.95	34.95	34.25	3912	488	488

2.3. Signal Conditioning

The experimental system captures five channels of the sEMG with a sampling frequency of 2000 Hz. A Butterworth bandpass filter (4~400 Hz) is applied to remove the offset and high frequency of the signal. A self-tuning notch filter is used to remove the power frequency of 50 Hz and its harmonics [36]. Audio is recorded with a sampling frequency of 16 kHz. One example of the collected signals with the audio and their five-channel sEMG signals is presented in Figure 2 and color is used to distinguish the 5 channels.
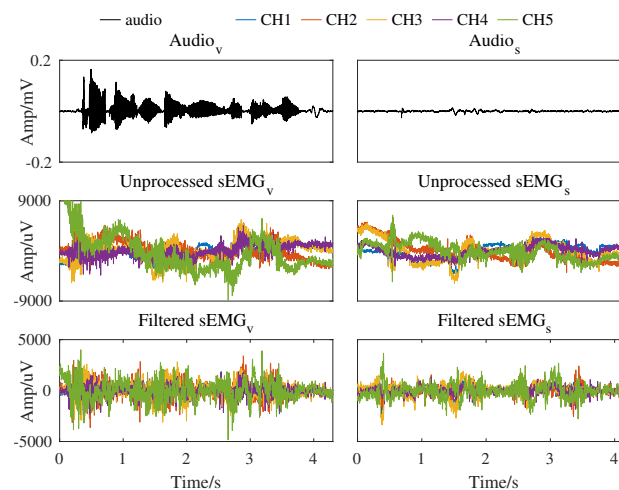
Figure 2. Time–series plots of the audio waveform and neuromuscular signal from Spk-3 with vocal mode and silent mode for the Chinese sentence "tui4 yi4 yun4 dong4 yuan2 you3 shen2 me5". The audio and neuromuscular signal of the same mode are collected time-synchronized.

2.4. Feature Extraction

To extract the feature of sEMG, we use time-domain (TD) features and time-frequency domain features from the amplitude of short-time Fourier transform (STFT), with a 64 ms Hanning window and 16 ms hop length [2,16]. Six TD features are calculated from one frame following [37]. Finally, 5×6-dimensional TD features and 5×65-dimensional STFT features are extracted and concentrated, i.e., 355-dimensional features are used as input to our model.

To maintain the alignment with sEMG, we extract an 80-dimensional mel-spectrogram with the band-limited frequency range (80~7600 Hz) from $Audio_v$, in which the window length is 1024 points and the hop length is 256 points [38].

3. The Proposed Methods

3.1. Overview

In order to distinguish between two kinds of sEMG modes, $X_{1:N}$ represents $sEMG_s$ features while $x_{1:M}$ represents $sEMG_v$ features. Additionally, $Y_{1:M}$ represents the mel-

spectrograms from $Audio_v$. The target task, i.e., the goal of the $sEMG2V$, is essentially to transform an N-length time-series sequence $X_{1:N}$ into an M-length sequence $Y_{1:M}$. Note that the length M of the target sequence $Y_{1:M}$ is unknown and depends on the source sequence itself.

To fulfill this task, we design a novel $sEMG2V$ model, called Silent Speech Reconstruction Network (SSRNet in short), see Figure 3. SSRNet generates the mel-spectrograms $Y_{1:M}$ directly from the features of $sEMG_s$ $X_{1:N}$. Moreover, SSRNet resamples the input sequence according to the duration sequence $d_{1:N}$ (i.e., $X_{1:N}[i]$, the index i of X_N corresponds to $[Y_{1:M}[j], \cdots, Y_{1:M}[j+p-1]]$ of Y_M, where p is the duration of $X_{1:N}[i]$, called $d_{1:N}[i]$) which is calculated from the alignment between $sEMG_s$ and $Audio_v$. Finally, SSRNet transfers the predicted mel-spectrograms $\hat{Y}_{1:M}^{+}$ to the audio waveform by a pre-trained vocoder.
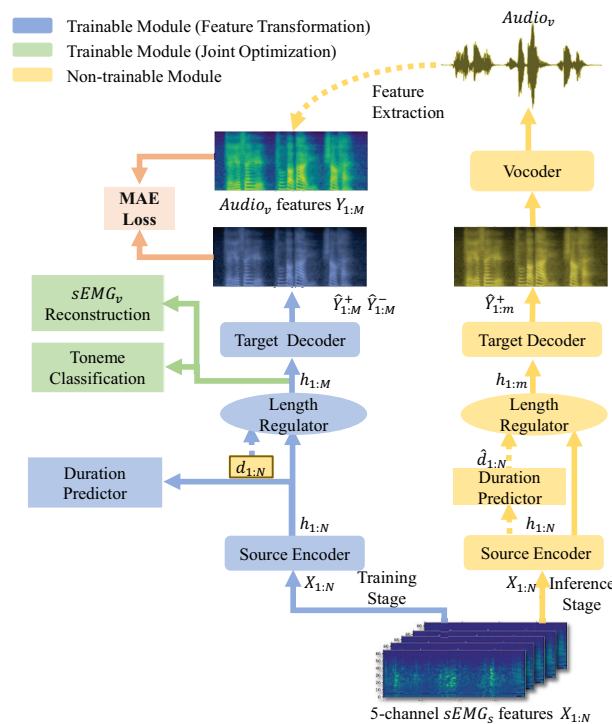
Figure 3. The overview of the training and inference stages in the SSRNet model. Blue and green blocks represent the feature transformation and joint optimization of the training module, respectively. Yellow blocks represent the non-trainable module, using a pre-trained model to predict the mel-spectrograms without the joint optimization part. We will detail the duration predictor in Section 3.3, and then detail the $sEMG_v$ reconstruction module and the toneme classification module in Section 3.4.

The procedure mentioned above can be formally described as follows:

$$h_{1:N} = \text{Encoder}(X_{1:N}) \tag{1}$$

where $h_{1:N}$ is the hidden representations produced by the source encoder.

$$h_{1:M} = \text{LengthRegulator}(h_{1:N}, d_{1:N}) \tag{2}$$

where $h_{1:M}$ is generated from $h_{1:N}$ by the length regulator, note that $M = \sum_{i=1}^{N} d_{1:N}[i]$ and $d_{1:N}$ is the ground-truth duration (GT duration) after the alignment.

$$\hat{Y}_{1:M}^{+} = \text{Decoder}(h_{1:M}) \tag{3}$$

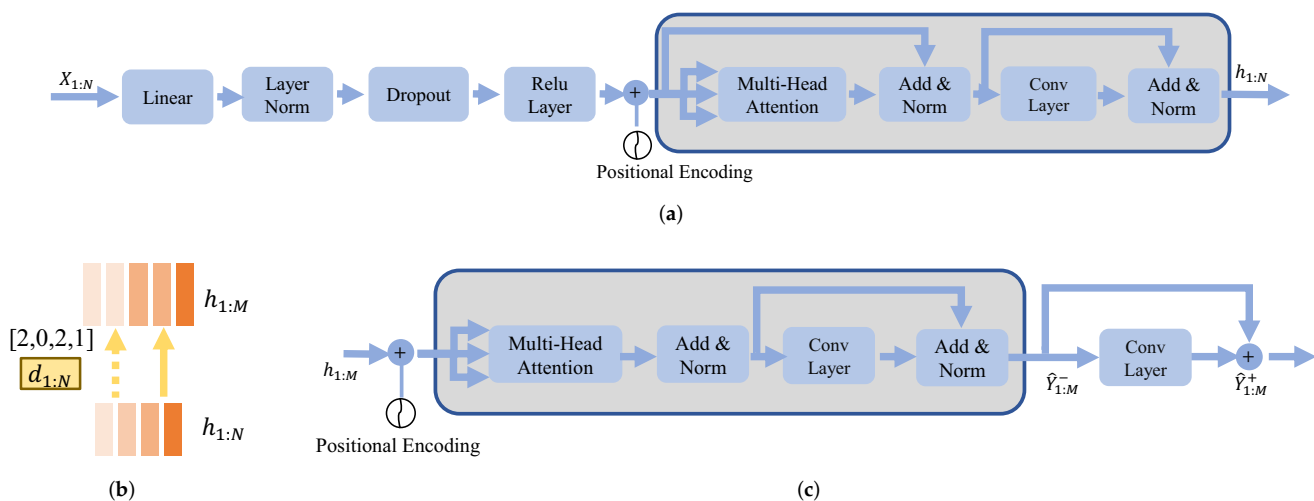where $\hat{Y}^{+}_{1:M}$ is the mel-spectrograms predicted by the decoder.

$$Audio = \text{Vocoder}\left(\hat{Y}^{+}_{1:M}\right). \tag{4}$$

In the inference stage, we use modules of the feature transformation and the duration predicted by the duration predictor instead of the GT duration. The inference stage is also illustrated in Figure 3, $h_{1:m}$ is the same as $h_{1:M}$; $\hat{d}_{1:N}$ is the predicted duration by the duration predictor; and $\hat{Y}^{+}_{1:m}$ represent the mel-spectrograms predicted in the inference module, where $m = \sum_{i=1}^{N} \hat{d}_{1:N}[i]$.

3.2. Feature Transformation

The feature transformation module aims to transform the sEMG features to audio features by the length regulator using GT duration. The architecture for the feature transformation in SSRNet includes an encoder, a length regulator, and a decoder. The main structure of SSRNet is called the feed-forward transformer (FFT) [39], which consists of self-attention in transformer and 1D convolution layers. The FFT aims at exploring the relationship between $X_{1:N}$ and $Y_{1:M}$ at different positions. This module follows the setting in [32].

The source encoder block, illustrated in Figure 4a, uses a fully connected layer with rectified linear unit (ReLU) activation to convert multi-dimension features of the sEMG to match the FFT hidden size [40]. The positional encoding is introduced to concatenate with the output of the linear layer in order to highlight the position of the frame in $X_{1:N}$. After that, SSRNet uses a multiple FFT structure (shown as gray blocks in Figure 4a,c) with multi-head attention and a two-layer 1D convolutional network.

Figure 4. Illustrations of model details about the source encoder, length regulator, and the target decoder. $\hat{Y}^{+}_{1:M}$ and $\hat{Y}^{-}_{1:M}$ are the mel-spectrograms predicted before and after the postnet. The gray blocks represent the FFT module. (**a**) Source encoder, (**b**) length regulator, (**c**) target decoder.

SSRNet applies a length regulator to adjust the length of output hidden representations of the source encoder block to match the output features. Figure 4b depicts the length regulator where the length of the input is four, while the length of the output is five. The length of the regulated sequence is adjusted to five by the GT duration $d_{1:N}$. The duration from the alignment between $X_{1:N}$ and $x_{1:M}$ is denoted as the GT duration which will be detailed in Section 3.3. Note that $d_{1:N}$ is only used in the training procedure. In the inference procedure, we use the output $\hat{d}_{1:N}$ from the duration predictor as duration to regulate.

The FFT layer used by the target decoder block is the same as the source encoder. As illustrated in Figure 4c, the output hidden representations after FFT blocks are passed

through the linear layer. Mel-spectrograms predicted after the linear layer are $\hat{Y}_{1:M}^{-}$. SSRNet further uses convolutional layers called postnet to calculate the residual of the predicted mel-spectrograms, which is used to improve the reconstruction ability of the model [41]. $\hat{Y}_{1:M}^{+}$ is the sum of $\hat{Y}_{1:M}^{-}$ and the residual mel-spectrograms.

In the feature transformation, SSRNet uses the mean absolute deviations error (MAE) as the loss function. To be more specific, we minimize the summed MAE of between $\hat{Y}_{1:M}^{+}$ and $Y_{1:M}$, and between $\hat{Y}_{1:M}^{-}$ and $Y_{1:M}$.
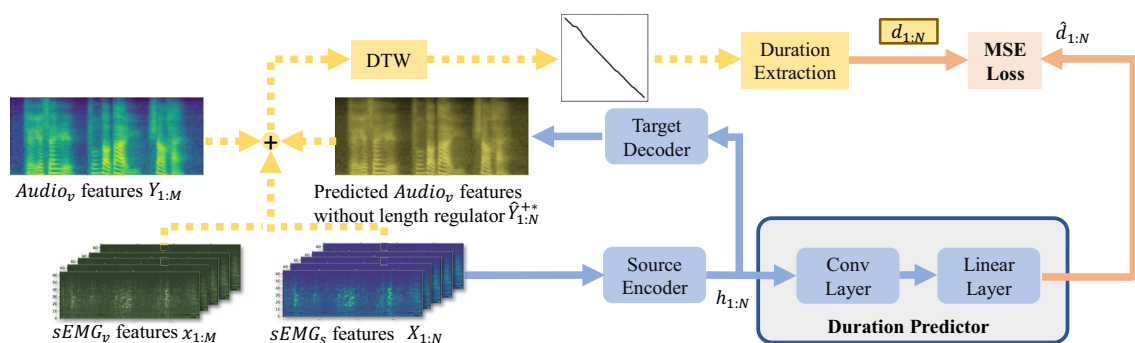
3.3. Duration Extractor

Given the synchronization between $X_{1:N}$ and $x_{1:M}$, the duration extractor uses dynamic programming to achieve pair positions between N-length $X_{1:N}$ and M-length $x_{1:M}$ [16,42]. The cost function is defined as follows:

$$\|X_{1:N}[i] - x_{1:M}[j]\|, \; 1 \le i \le N, 1 \le j \le M. \tag{5}$$

In addition, similar to the predicted audio refinement [16], the model without the length regulator obtains N-length predicted audio features $\hat{Y}_{1:N}^{+*}$ during the training procedure. As illustrated in Figure 5, the new cost function for DTW in this method is shown as follows:

$$\|X_{1:N}[i] - x_{1:M}[j]\| + \lambda_{align}\|\hat{Y}_{1:N}^{+*}[i] - Y_{1:M}[j]\| \tag{6}$$

where λ_{align} is the weight of audio alignments.

Figure 5. Illustration of duration extraction and predictor. The gray block represents the duration predictor.

Instead of achieving a warped audio sequence from pairs, the proposed SSRNet model calculates the duration sequence from the pairs as follows:

$$d_{1:N}[i] = \sum_{j=1}^{M}(A_{1:M}[j] == i) \tag{7}$$

where $A_{1:M}$ is the length of M sequence which represents whether the index i of the input features is corresponding to the index j of the output features.

The duration predictor aims at predicting the length of audio features corresponding to each frame of sEMG features. The duration extractor is based on the DTW algorithms [30]. SSRNet trains a duration predictor (i.e., convolutional layers and a linear layer) and uses mean square error (MSE) to calculate the loss between GT duration $d_{1:N}$ and the predicted duration $\hat{d}_{1:N}$.

3.4. Joint Optimization with Toneme Prediction and Vocal sEMG Reconstuction

The module of joint optimization with toneme prediction and vocal sEMG reconstruction aims at improving the model performance. SSRNet employs the pre-trained Mandarin model from Montreal Forced Aligner (MFA) to obtain the toneme alignment $tm_{1:M}$ of the

audio [43,44]. The set of tonemes for Mandarin is created by GlobalPhone [28] by splitting into onset, nucleus (any vowel sequence), and codas, and then associating the tone of the syllable onto the nucleus (i.e., /teng2/ is split as /t e2 ng/). In Figure 6, the hidden representations pass to a linear layer to predict a sequence (including silent frames) $\hat{tm}_{1:M}$, and SSRNet uses the cross-entropy (CE) to measure the loss between the target and the output. The purpose of the module is to conserve information of the target context.
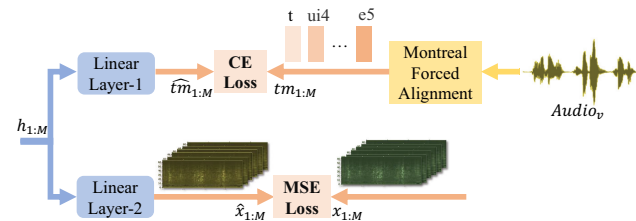
Figure 6. Detail of joint optimization. $tm_{1:M}$ is the tonemes with tones from MFA while $\hat{tm}_{1:M}$ is the tonemes predicted after the linear layer. $x_{1:M}$ is the $sEMG_v$ while $\hat{x}_{1:M}$ is the $sEMG_v$ reconstructed from the linear layer.

In addition, another linear layer at the same position in Figure 6 is used to restore the hidden representation to the $sEMG_v$ for the stable training procedure.

During the inference stage, the joint optimization module is discarded. The joint loss function of the proposed SSRNet model is formulated as follows:

$$
\begin{aligned}
\mathcal{L}_{SSRNet} = {} & MAE(\hat{Y}^{+}_{1:M}, Y_{1:M}) \\
& + MAE(\hat{Y}^{-}_{1:M}, Y_{1:M}) \\
& + MSE(\hat{d}_{1:N}, d_{1:N}) \\
& + \lambda_{tm} CE(\hat{tm}_{1:M}, tm_{1:M}) \\
& + \lambda_{recons} MSE(\hat{x}_{1:M}, x_{1:M})
\end{aligned}
\tag{8}
$$

where λ_{tm} controls the toneme classification loss and λ_{recons} controls the $sEMG_v$ reconstruction loss.

3.5. Vocoder

This paper utilizes Parallel WaveGAN (PWG) as the final synthesizer of desired audible speech [38]. This vocoder is an upgraded non-autoregressive version of the WaveNet model [45]. Unlike some previous non-autoregressive methods such as [46–48], PWG gets rid of the teacher–student framework, which significantly facilitates our training process and speeds up in the inference stage.

To synthesize natural $Audio_v$, PWG requires an input of auxiliary features, which is $Y_{1:M}$ for training and $\hat{Y}_{1:M}$ for inference. The model consists of a non-autoregressive WaveNet generator and a discriminator with non-causal dilated convolution. Instead of the traditional sequential teacher–student framework, PWG has a structure of a generative adversarial network (GAN) and jointly optimizes adversarial function loss L_{adv} and the auxiliary loss L_{aux} of multi-resolution STFT loss [45]. The loss function of the multi-tasking generator is defined as:

$$
\begin{aligned}
L_G(G, D) = {} & \lambda_{adv} L_{adv}(G, D) \\
& + E_{v \sim p_{data}, z} L_{aux}(v, \hat{v})
\end{aligned}
\tag{9}
$$

where v is the original audio while $\hat{v} = G(z, Y_{1:M})$ is the predicted audio, p_{data} represents the distribution of ground-truth waveform data, z represents our injected Gaussian noise, and λ_{adv} is a tunable parameter to balance the performance between tasks.

On the other hand, loss equation of the discriminator defined below aims at strengthening its ability to tell the generated waveforms from the ground-truth:

$$L_D(G, D) = E_{v \sim p_{data}}\left[(1 - D(v))^2\right]$$
$$+ E_z\left[D(G(z, Y_{1:M}))^2\right]. \tag{10}$$

The block diagram of PWG is shown in Figure 7. The generator and discriminator are optimized according to a certain strategy during the training stage, and the trained generator is further used in the inference stage to produce the final results of the SSRNet.
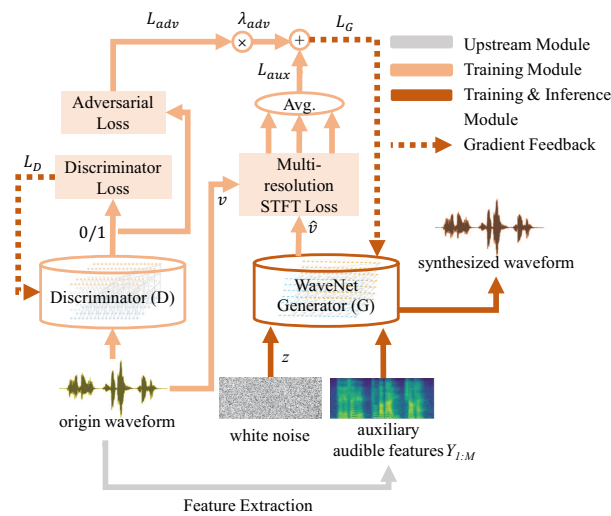
Figure 7. The framework of PWG. Orange blocks represent the training module, while red blocks represent both the training and inference modules. Dotted lines mean gradient feedback.

4. Experiments and Results

4.1. Experimental Setting

In the training stage of the SSRNet, the batch size is set to 8 utterances. In addition, the dropout rate for encoder and decoder is set to 0.1 and for postnet it is set to 0.5. The detailed settings of SSRNet are shown in Table 3. The Adam optimization algorithm is used to optimize trainable parameters. The Noam learning rate (LR) scheduler is used in the training procedure as follows [39]:

$$lr = d_{model}^{-0.5} * \min(step^{-0.5}, step * step_w^{-1.5}) \tag{11}$$

where $step_w$ is set to 4000, d_{model} is set to 384, and $step$ denotes the number of the training steps. These parameter values are chosen based on [39]. Furthermore, λ_{align} in Equation (6) is set to 10, λ_{tm} and λ_{recons} in Equation (8) are both set to 0.5. The GT duration of the training set is calculated as Equations (5) and (7) before training. The model uses this initial GT duration to calculate the loss in the first four epochs. In the training stage, the GT duration is updated every five epochs by Equations (6) and (7). The implementation of the SSRNet model is based on the ESPNET toolkit (https://github.com/espnet/espnet, accessed on 22 March 2022) [49].

Table 3. Hyper-parameters of SSRNet.

Item		Details
attention transformation dimensions		384
heads for multi-head attention		4
source encoder	FFT layers	6
	hidden units	1536
target decoder	FFT layers	6
	hidden units	1536
postnet	layers	5
	filter channels	256
	filter size	5
duration predictor	layers	2
	filter channels	384
	kernel size	3

For the vocoder, PWG is pre-trained within $Audio_v$ of multi-speakers in the training set. In the first 100 K steps of the training stage, the discriminator parameters are fixed, and only the generator is trained on the first stage. After that, the two modules are jointly trained until 400 K steps to further build the synthesis quality. Our experiment is based on the implementation of PWG (https://github.com/kan-bayashi/ParallelWaveGAN, accessed on 22 March 2022). The detailed settings of PWG are shown in Table 4.

Table 4. Hyper-parameters of PWG.

Item	Objects and/or Details	
λ_{adv}	4	
filter size	3	
batch size	6	
training audio length	16,384 (1.024 s)	
WaveNet generator	30-layer dilated residual convolution	
discriminator	10-layer dilated residual convolution	
learning rate	generator	1×10^{-4}
	discriminator	5×10^{-5}
training steps	generator-only	1×10^6
	jointly	4×10^6
channel size	skip channels	64
	residual channels	64
optimizer	RAdam optimizer	$\epsilon = 1 \times 10^{-6}$
activation function	Leaky ReLU	$\alpha = 0.2$

The previous work proposed by Gaddy and Klevin [16] is considered as the baseline model. The training parameters of the baseline model are consistent with those reported in [16]. The training, validation, and testing data used are the same as those used in the SSRNet model. Moreover, we employ the pre-trained PWG instead of WaveNet as the vocoder in baseline to deal with the limitation of inference speed [38]. We train the SSRNet and baseline separately for each participant.

4.2. Model Performance on the sEMG_Mandarin Dataset

4.2.1. Objective Evaluation

The objective evaluation is about the quality and accuracy of reconstructed voices. For the objective accuracy evaluation, this paper employs an automatic speech recognition (ASR), called Mandarin ASR (MASR) (https://github.com/nobody132/masr, accessed on 22 March 2022), as a metric. MASR uses the character error rate (CER) with the Levenshtein distance to measure the accuracy between the predicted text and the original text [50]. Note that CER ranges from 0 to +∞. CER can become infinite because the ASR can insert an arbitrary number of words [51]. In the experiments, CER based on ASR for each epoch is calculated on the validation set by the model, and parameters of the best CER epoch are selected as the best-performing final model.

It is observed in Figure 8 that the proposed method SSRNet outperforms the baseline significantly for all six speakers. The SSRNet obtains an average CER of 21.99% in ASR with a standard deviation of 4.99% across six speakers. In addition, the SSRNet outperforms the baseline by 24.63%. Meanwhile, the ground-truth voices from the testing set achieve a CER of 11.30%. This verifies that SSRNet generates more intelligible voices. This occurs because SSRNet calculates the duration of the silent speech, regulates the silent sEMG following audio length, and uses a multi-task learning strategy to improve results. Additionally, the results across speakers differ, among which the worst accuracy is achieved on Spk-4 with a CER of 27.20% and Spk-5 with a CER of 27.34%; the best accuracy is achieved on Spk-1 with a CER of 13.62%. By studying the two speaker cases with the worst accuracy, we find that the ground-truth voices of Spk-4, which performs poorly on ASR, can cause low accuracy. At the same time, higher impedance resulting in a lower signal-to-noise ratio during the experiment leads to the wrong result on Spk-5.
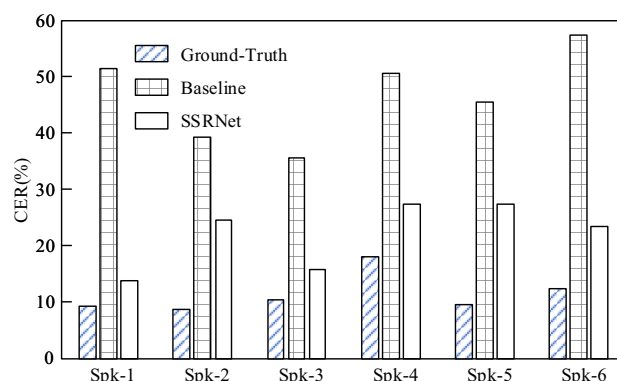
Figure 8. Objective accuracy comparison between the ground-truth voices, reconstructed voices from the baseline, and SSRNet.

For the objective quality evaluation, we use mel-cepstral distortion (MCD) (https://github.com/mpariente/pystoi, accessed on 22 March 2022) [52] and short-term objective intelligibility (STOI) (https://github.com/ttslr/python-MCD, accessed on 22 March 2022) [53]. The lower MCD indicates a higher similarity between the synthesized and the natural mel-cepstral sequences. Meanwhile, the higher STOI reflects higher intelligibility and better clarity of the speech.

Figure 9 summarizes the MCD and STOI evaluation. It is observed that the SSRNet model consistently performs better than the baseline model for both quality and intelligibility. The reason is that the length of reconstructed voice in the baseline is consistent with silent speech and impaired. As a comparison, SSRNet firstly provides length-regulated voices, which are more similar to the ground-truth voices.
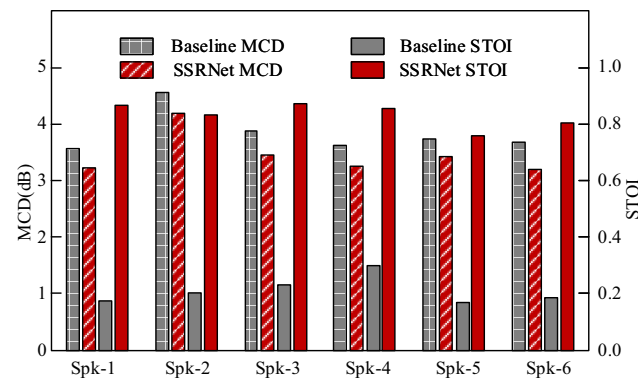
Figure 9. Objective quality comparison between reconstructed voices from the baseline and SSRNet.

4.2.2. Subjective Evaluation

We use subjective evaluation based on the transcriptions from 10 native Mandarin Chinese human listeners. The average age of the ten listeners is 24. These listeners have no prior knowledge of the context of the voices. They are required to listen to the voices with earphones in a quiet environment. Each listener is required to listen to 60 sample voices from 6 speakers, randomly selected from SSRNet and baseline testing set. They are asked to transcribe the audios into text in Mandarin Chinese and give a score of the naturalness of each speaker ranging from 0 to 100 (0 for the worst naturalness while 100 for the best).

The reconstructed samples can be found on our website (https://irislhy.github.io/, accessed on 22 May 2022).

The results of human evaluation of six speakers' samples are shown in Table 5 and ± indicates the standard deviation of the metrics across the evaluation of the listeners. The results of the subjective evaluation are consistent with the objective evaluation. SSRNet obtains an average CER of 6.41%, while the baseline obtains an average CER of 39.76% in subjective evaluation. Additionally, the naturalness scores from listeners are consistent with the objective evaluation results. Our exploratory analysis shows that the proposed SSRNet outperforms the baseline in human intelligibility and naturalness.

Table 5. Subjective Comparison between Reconstructed Voices from the Baseline and SSRNet.

	Spk-1	Spk-2	Spk-3	Spk-4	Spk-5	Spk-6
Baseline CER (%)	55.06 ± 41.62	17.72 ± 13.96	23.00 ± 2.71	53.37 ± 6.67	26.37 ± 21.18	63.05 ± 27.00
SSRNet CER (%)	1.70 ± 3.4	1.19 ± 1.46	2.31 ± 2.37	8.92 ± 5.77	20.67 ± 5.69	3.69 ± 3.05
Baseline Naturalness	44 ± 16	50 ± 14	51 ± 17	39 ± 13	51 ± 20	41 ± 18
SSRNet Naturalness	95 ± 7	71 ± 17	89 ± 5	64 ± 10	58 ± 11	77 ± 16

4.2.3. Mel-Spectrogram Comparison

Figure 10 depicts mel-spectrograms of one example from the testing set of Spk-3. We have three observations: (1) The mel-spectrograms synthesized by SSRNet are closer to the ground-truth one and have a similar length. This is because the length regulator resamples the length of the output frames. Based on this, SSRNet deals with the lack of time-aligned data of vocal and silent speech, and generates more natural sounds. (2) The white box indicates a slight blurring of the pronunciation of /ye3/ in SSRNet compared to the ground-truth pronunciation, but listeners and ASR can understand it. (3) The yellow boxes indicate four examples of errors for the baseline that some listeners and ASR have difficulty understanding, and the voice synthesized from the baseline is not clear overall.
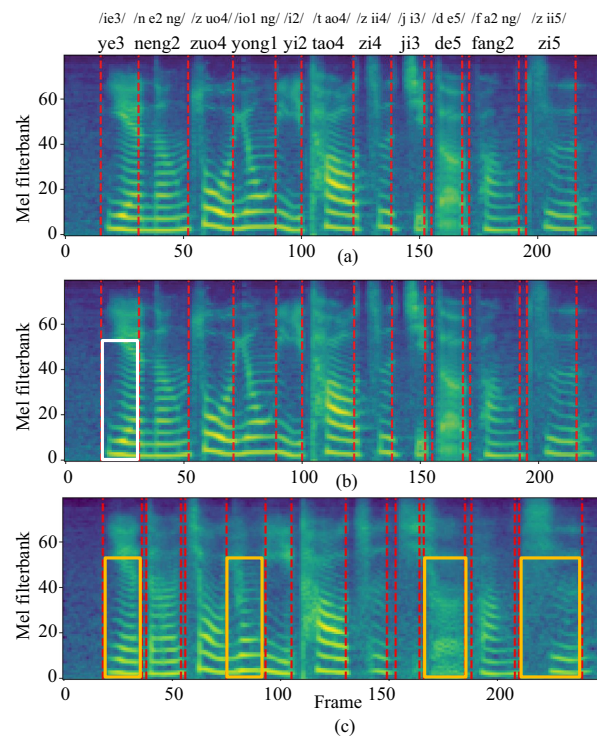
Figure 10. Mel-spectrogram visualizations of (**a**) ground-truth voice, (**b**) voice reconstructed by SSRNet, (**c**) voice reconstructed by the baseline.

In conclusion, the experiments demonstrate that SSRNet provides a solution to narrow the gap between the reconstructed and natural voices.

4.3. Ablation Study

Next, we conduct ablation studies to gauge the effectiveness of every extension in SSRNet, including joint optimization, model prediction alignments, and tone evaluation. Due to the consistency between the objective and subjective evaluation, only the objective accuracy evaluation is performed for ablation studies. Table 6 summarizes the ablation study results of different model modules. The first row shows the settings of SSRNet, while the final column shows the change in average CER across six speakers compared to SSRNet.

Table 6. Ablation Results from the Model Module Study.

$sEMG_v$ Reconstruction Module	Toneme Classification Module	Toneme Classification Module Position	Tones in Toneme Classification Module	Cost Function for DTW	ΔCER (%)
✓ ($\lambda_{recons} = 0.5$)	✓ ($\lambda_{tm} = 0.5$)	Before Decoder	✓	$\lambda_{align} = 10$	+0
✗ ($\lambda_{recons} = 0$)	✓ ($\lambda_{tm} = 0.5$)	Before Decoder	✓	$\lambda_{align} = 10$	+9.38
✓ ($\lambda_{recons} = 0.5$)	**✗ ($\lambda_{tm} = 0$)**	-	✓	$\lambda_{align} = 10$	+132.75
✓ ($\lambda_{recons} = 0.5$)	✓ ($\lambda_{tm} = 0.5$)	**After Decoder**	✓	$\lambda_{align} = 10$	+1.89
✓ ($\lambda_{recons} = 0.5$)	✓ ($\lambda_{tm} = 0.5$)	Before Decoder	✗	$\lambda_{align} = 10$	+6.51
✓ ($\lambda_{recons} = 0.5$)	✓ ($\lambda_{tm} = 0.5$)	Before Decoder	✓	**$\lambda_{align} = 0$**	+81.18

4.3.1. Joint Optimization

The second and third rows show the changes in consequences of removing the joint optimization. It is observed that removing the joint optimization could lead to performance degradation in terms of accuracy. This indicates that the toneme classification and

the $sEMG_v$ reconstruction are practical for SSRNet. Note that the toneme classification module contributes significantly more in SSRNet than $sEMG_v$ reconstruction. We find that removing the toneme classification results in an absolute difference between the context of synthesized voices and the ground-truth context. It means that in the Seq2Seq model, the hidden representations after the length regulator have difficulty in obtaining the context information of the $sEMG_s$. As a result, joint optimization is conducive to studying feature transformation.

4.3.2. The Position of the Toneme Classification Module

We also investigate the position of the toneme classification module by comparing results in the fourth row with those in the first row while the position of $sEMG_v$ reconstruction is fixed. In the fourth row of the table, the position of the toneme classification is located after the decoder. In contrast, the position in the first row is located before the decoder. The position before the decoder outperforms the position after the decoder by 1.89%. This implies that the position of the module in the middle layer or final layer can both represent the source content in the $sEMG2V$ task.

4.3.3. Tone in Toneme Classification

We also conduct the tone evaluation. We use phoneme classification instead of toneme classification. The phoneme classification module predicts a sequence and measures the CE loss between true and predicted phonemes without any tone information. We find that lack of tone information resulted in a 6.51% increase in CER in the fifth row, which demonstrates that the $sEMG2V$ task in Mandarin Chinese needs tone information in concert with phoneme rather than separate phoneme information.

4.3.4. Cost Function for DTW

We conduct the alignment study as described in the sixth row. It shows that the CER of the alignment strategy in SSRNet shows a relative reduction of over 81.18% compared to the traditional approach, which demonstrates that effectiveness of the alignments between N-length predicted audio features $\hat{Y}_{1:N}^{+*}$ obtained by the SSRNet model without a length regulator and M-length ground-truth audio features $Y_{1:M}$.

4.4. Frame-Based Toneme Classification Study

Finally, we evaluate the frame-based performance of the toneme classification module on the testing set except for silent frames. We use the GT-duration calculated by Equation (7) with the best-performing model of each speaker to match the length of ground-truth phonemes. As the confusion between vowel consonants is interpretable [54], this section focuses on the vowel pairs, consonants pairs, and tone pairs. The confusion matrices are calculated to elaborate more toneme prediction details, as shown in Figure 11.

It can be seen in Figure 11a,b that SSRNet provides excellent classification results for consonants and vowels. We observe the confusion between nasal and other consonants, which is consistent with [54,55]. This is due to the limitations of sEMG electrodes in detecting velum [55].

Meanwhile, Figure 11c shows the confusion matrix of the tone set, which is calculated from the ground-truth tones and the predicted tones from vowels and is directly extracted from the entire confusion matrix. The tone classification achieves an average accuracy of 96.07%. This proves that neuromuscular signals can transfer most tone information in silent speech. The fifth tone is sometimes mistaken for the other four tones. This indicates that the fifth tone is sometimes difficult to express in silent speech.

(a)

(b)

(c)
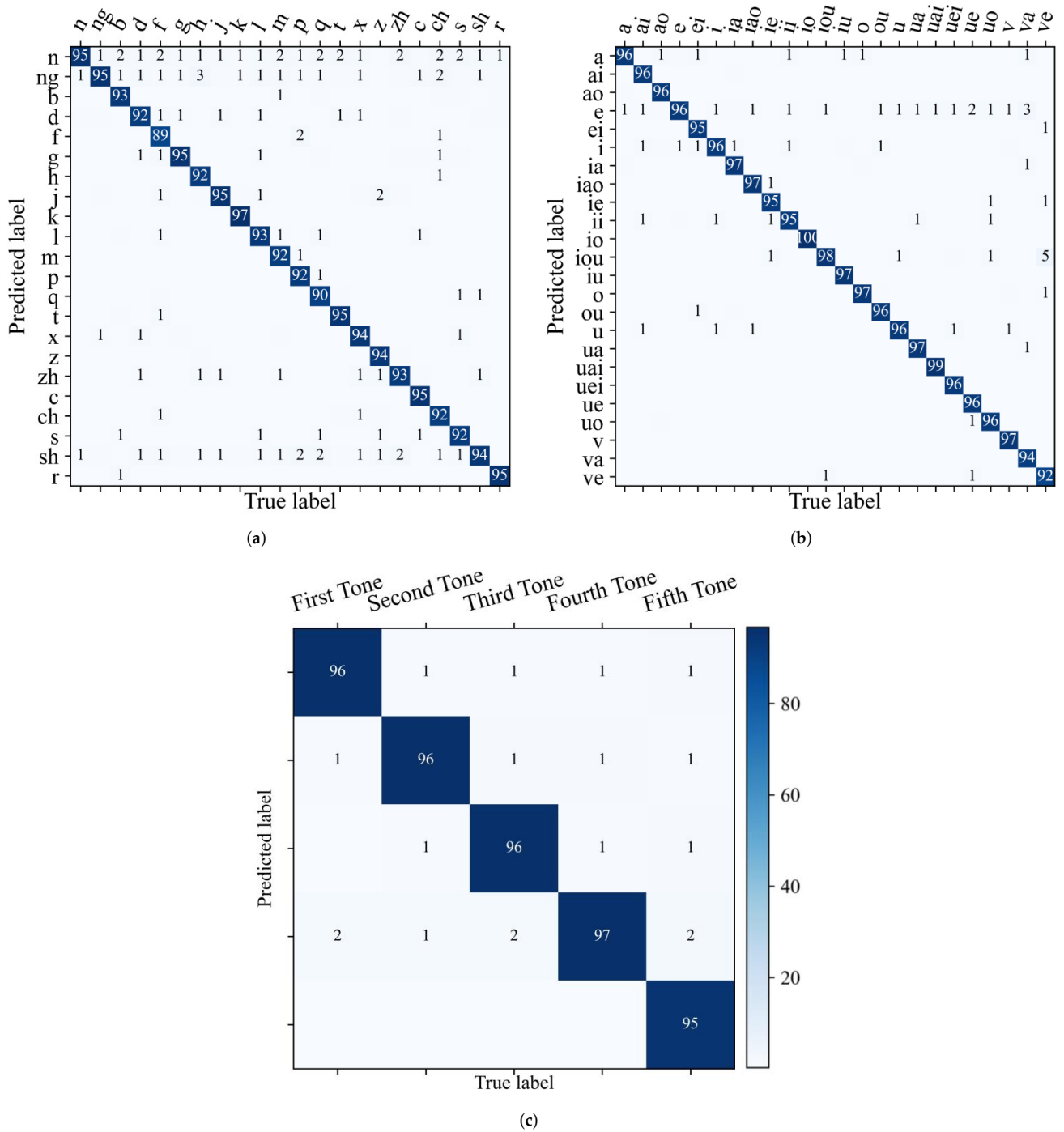
Figure 11. Toneme confusion matrices on the testing set, with the number in row j, column i is the ratio in percentage of the number of samples predicted as label j with the true label i to the number of samples with the true label i. Values smaller than 0.5% are ignored. (**a**) Consonant confusability, (**b**) vowel confusability, (**c**) tone confusability.

5. Conclusions

This paper proposes a Seq2Seq-based SSRNet model to decode neuromuscular signals in a tonal language. SSRNet uses the duration extracted from the alignment to regulate the sEMG-based silent speech. Furthermore, a toneme classification module and a vocal sEMG reconstruction module are used to improve the overall performance. We conduct extensive experiments on a Mandarin Chinese dataset to demonstrate that the proposed model

outperforms the baseline model in both objective and subjective evaluation. The model achieves an average subjective CER of 6.41% for six speakers and 1.19% for the best speaker, demonstrating the feasibility of the reconstruction task.

In the future, we would like to enhance the robustness and generalization of the model by including more speakers and utilizing transfer learning. Another possible direction is making the system real-time because it is necessary for speakers to learn to improve pronunciation by themselves in silent speech based on auditory feedback.

Author Contributions: Conceptualization, Y.W. and G.L.; Data curation, H.L. (Huiyan Li), H.G. and Q.A.; Formal analysis, H.L. (Huiyan Li), H.L. (Haohong Lin) and Q.A.; Investigation, H.L. (Huiyan Li), H.W. and H.G.; Methodology, H.L. (Huiyan Li), H.L. (Haohong Lin) and Y.W.; Resources, H.L. (Huiyan Li); Software, H.L. (Huiyan Li), H.W. and M.Z.; Validation, Y.W.; Visualization, H.L. (Huiyan Li), H.L. (Haohong Lin) and H.W.; Writing—original draft, H.L. (Huiyan Li) and H.L. (Haohong Lin); Writing—review and editing, Y.W., M.Z. and Z.L. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the College of Biomedical Engineering and Instrument Science Ethical Committee of Zhejiang University (Ethical Approval: ZJU IRB CBEIS-2022[025]).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Denby, B.; Schultz, T.; Honda, K.; Hueber, T.; Gilbert, J.M.; Brumberg, J.S. Silent speech interfaces. *Speech Commun.* **2010**, *52*, 270–287. [CrossRef]
2. Wang, Y.; Zhang, M.; Wu, R.; Gao, H.; Yang, M.; Luo, Z.; Li, G. Silent speech decoding using spectrogram features based on neuromuscular activities. *Brain Sci.* **2020**, *10*, 442. [CrossRef] [PubMed]
3. Schultz, T.; Wand, M.; Hueber, T.; Krusienski, D.J.; Herff, C.; Brumberg, J.S. Biosignal-Based Spoken Communication: A Survey. *IEEE ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 2257–2271. [CrossRef]
4. López, J.A.G.; Alanís, A.G.; Martín-Doñas, J.M.; Pérez-Córdoba, J.L.; Gomez, A.M. Silent Speech Interfaces for Speech Restoration: A Review. *IEEE Access* **2020**, *8*, 177995–178021. [CrossRef]
5. Herff, C.; Heger, D.; de Pesters, A.; Telaar, D.; Brunner, P.; Schalk, G.; Schultz, T. Brain-to-text: Decoding spoken phrases from phone representations in the brain. *Front. Neurosci.* **2015**, *9*, 217. [CrossRef]
6. Angrick, M.; Herff, C.; Johnson, G.D.; Shih, J.J.; Krusienski, D.J.; Schultz, T. Interpretation of convolutional neural networks for speech spectrogram regression from intracranial recordings. *Neurocomputing* **2019**, *342*, 145–151. [CrossRef]
7. Angrick, M.; Herff, C.; Mugler, E.; Tate, M.C.; Slutzky, M.W.; Krusienski, D.J.; Schultz, T. Speech synthesis from ECoG using densely connected 3D convolutional neural networks. *J. Neural Eng.* **2019**, *16*, 036019. [CrossRef]
8. Ramadan, R.A.; Vasilakos, A.V. Brain computer interface: Control signals review. *Neurocomputing* **2017**, *223*, 26–44. [CrossRef]
9. Porbadnigk, A.; Wester, M.; Calliess, J.; Schultz, T. EEG-based Speech Recognition—Impact of Temporal Effects. In Proceedings of the International Conference on Bio-inspired Systems and Signal Processing, Porto, Portugal, 14–17 January 2009; pp. 376–381.
10. Rolston, J.D.; Englot, D.J.; Cornes, S.; Chang, E.F. Major and minor complications in extraoperative electrocorticography: A review of a national database. *Epilepsy Res.* **2016**, *122*, 26–29. [CrossRef]
11. Diener, L.; Janke, M.; Schultz, T. Direct conversion from facial myoelectric signals to speech using Deep Neural Networks. In Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 12–17 July 2015; pp. 1–7.
12. Wand, M.; Janke, M.; Schultz, T. Tackling Speaking Mode Varieties in EMG-Based Speech Recognition. *IEEE Trans. Biomed. Eng.* **2014**, *61*, 2515–2526. [CrossRef]
13. Fagan, M.J.; Ell, S.R.; Gilbert, J.M.; Sarrazin, E.; Chapman, P.M. Development of a (silent) speech recognition system for patients following laryngectomy. *Med. Eng. Phys.* **2008**, *30*, 419–425. [CrossRef] [PubMed]
14. Denby, B.; Stone, M. Speech synthesis from real time ultrasound images of the tongue. In Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, 17–21 May 2004; pp. 685–688.

15. Janke, M.; Diener, L. EMG-to-Speech: Direct Generation of Speech From Facial Electromyographic Signals. *IEEE ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 2375–2385. [CrossRef]

16. Gaddy, D.; Klein, D. Digital Voicing of Silent Speech. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 5521–5530.

17. Meltzner, G.S.; Heaton, J.T.; Deng, Y.; Luca, G.D.; Roy, S.H.; Kline, J.C. Silent Speech Recognition as an Alternative Communication Device for Persons With Laryngectomy. *IEEE ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 2386–2398. [CrossRef] [PubMed]

18. Zhang, M.; Wang, Y.; Zhang, W.; Yang, M.; Luo, Z.; Li, G. Inductive conformal prediction for silent speech recognition. *J. Neural Eng.* **2020**, *17*, 066019. [CrossRef]

19. Wang, Y.; Zhang, M.; Wu, R.; Wang, H.; Luo, Z.; Li, G. Speech neuromuscular decoding based on spectrogram images using conformal predictors with Bi-LSTM. *Neurocomputing* **2021**, *451*, 25–34. [CrossRef]

20. Li, Y.; Tang, C.; Lu, J.; Wu, J.; Chang, E.F. Human cortical encoding of pitch in tonal and non-tonal languages. *Nat. Commun.* **2021**, *12*, 1–12. [CrossRef]

21. Kaan, E.; Wayland, R.; Keil, A. Changes in oscillatory brain networks after lexical tone training. *Brain Sci.* **2013**, *3*, 757–780. [CrossRef]

22. Huang, W.; Wong, L.L.; Chen, F. Just-Noticeable Differences of Fundamental Frequency Change in Mandarin-Speaking Children with Cochlear Implants. *Brain Sci.* **2022**, *12*, 443. [CrossRef]

23. Chen, Y.; Gao, Y.; Xu, Y. Computational Modelling of Tone Perception Based on Direct Processing of f 0 Contours. *Brain Sci.* **2022**, *12*, 337. [CrossRef]

24. Surendran, D.; Levow, G.; Xu, Y. Tone Recognition in Mandarin Using Focus. In Proceedings of the 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005; pp. 3301–3304.

25. Yip, M. *Tone*; Cambridge University Press: Cambridge, UK, 2002.

26. Lei, X.; Ji, G.; Ng, T.; Bilmes, J.A.; Ostendorf, M. DBN-Based Multi-stream Models for Mandarin Toneme Recognition. In Proceedings of the (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, USA, 23 March 2005; pp. 349–352.

27. Trask, R.L. *A Dictionary of Phonetics and Phonology*; Routledge: London, UK, 2004.

28. Schultz, T.; Schlippe, T. GlobalPhone: Pronunciation Dictionaries in 20 Languages. In Proceedings of the Ninth International Conference on Language Resources and Evaluation, Reykjavik, Iceland, 26–31 May 2014; pp. 337–341.

29. Roach, P.; Widdowson, H. *Phonetics*; Oxford University Press: Oxford, UK, 2001.

30. Berndt, D.J.; Clifford, J. Using Dynamic Time Warping to Find Patterns in Time Series. In Proceedings of the AAAIWS'94: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, Online, 31 May 1994; pp. 359–370.

31. Hayashi, T.; Huang, W.; Kobayashi, K.; Toda, T. Non-Autoregressive Sequence-To-Sequence Voice Conversion. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Online, 6–11 June 2021; pp. 7068–7072.

32. Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T. FastSpeech: Fast, Robust and Controllable Text to Speech. In Proceedings of the Annual Conference on Neural Information Processing Systems 2019, Vancouver, BC, Canada, 8–14 December 2019; pp. 3165–3174.

33. Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.

34. Kim, J.; Kim, S.; Kong, J.; Yoon, S. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. In Proceedings of the Annual Conference on Neural Information Processing Systems, Online, 6–12 December 2020.

35. Shi, Y.; Bu, H.; Xu, X.; Zhang, S.; Li, M. AISHELL-3: A Multi-speaker Mandarin TTS Corpus and the Baselines. *arXiv* **2020**, arXiv:2010.11567.

36. Benesty, J.; Makino, S.; Chen, J. *Speech Enhancement*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006.

37. Jou, S.C.; Schultz, T.; Walliczek, M.; Kraft, F.; Waibel, A. Towards continuous speech recognition using surface electromyography. In Proceedings of the Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 17–21 September 2006.

38. Yamamoto, R.; Song, E.; Kim, J. Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Barcelona, Spain, 4–8 May 2020; pp. 6199–6203.

39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.

40. Huang, W.; Hayashi, T.; Wu, Y.; Kameoka, H.; Toda, T. Voice Transformer Network: Sequence-to-Sequence Voice Conversion Using Transformer with Text-to-Speech Pretraining. In Proceedings of the 21st Annual Conference of the International Speech Communication Association, Shanghai, China, 25–29 October 2020; pp. 4676–4680.

41. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Ryan, R.; et al. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing, Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783.

42. Desai, S.; Raghavendra, E.V.; Yegnanarayana, B.; Black, A.W.; Prahallad, K. Voice conversion using Artificial Neural Networks. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing, Taipei, Taiwan, 19–24 April 2009; pp. 3893–3896.

43. McAuliffe, M.; Socolof, M.; Mihuc, S.; Wagner, M.; Sonderegger, M. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In Proceedings of the 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, 20–24 August 2017; pp. 498–502.

44. Wang, D.; Zhang, X. THCHS-30: A Free Chinese Speech Corpus. *arXiv* **2015**, arXiv:1512.01882.

45. van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.W.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. In Proceedings of the 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13–15 September 2016; p. 125.

46. Prenger, R.; Valle, R.; Catanzaro, B. Waveglow: A flow-based generative network for speech synthesis. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12–17 May 2019; pp. 3617–3621.

47. Peng, K.; Ping, W.; Song, Z.; Zhao, K. Non-Autoregressive Neural Text-to-Speech. In Proceedings of the 37th International Conference on Machine Learning, Online, 12–18 July 2020; Volume 119; pp. 7586–7598.

48. Song, E.; Byun, K.; Kang, H.G. ExcitNet vocoder: A neural excitation model for parametric speech synthesis systems. In Proceedings of the 2019 27th European Signal Processing Conference (EUSIPCO), La Coruña, Spain, 2–6 September 2019, pp. 1–5.

49. Watanabe, S.; Hori, T.; Karita, S.; Hayashi, T.; Nishitoba, J.; Unno, Y.; Soplin, N.E.Y.; Heymann, J.; Wiesner, M.; Chen, N.; et al. ESPnet: End-to-End Speech Processing Toolkit. In Proceedings of the 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2–6 September 2018; pp. 2207–2211.

50. Navarro, G. A guided tour to approximate string matching. *ACM Comput. Surv.* **2001**, *33*, 31–88. [CrossRef]

51. Errattahi, R.; Hannani, A.E.; Ouahmane, H. Automatic Speech Recognition Errors Detection and Correction: A Review. *Procedia Comput. Sci.* **2018**, *128*, 32–37. [CrossRef]

52. Kubichek, R. Mel-cepstral distance measure for objective speech quality assessment. In Proceedings of the IEEE Pacific Rim Conference on Communications Computers and Signal Processing, Victoria, BC, Canada, 19–21 May 1993; Volume 1, pp. 125–128.

53. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In Proceedings of the 35th International Conference on Acoustics, Speech, and Signal Processing, Dallas, TX, USA, 14–19 May 2010; pp. 4214–4217.

54. Gaddy, D.; Klein, D. An Improved Model for Voicing Silent Speech. In Proceedings of the Association for Computational Linguistics (ACL) and Asian Federation of Natural Language Processing (AFNLP), Bangkok, Thailand, 1–6 August 2021; pp. 175–181.

55. Freitas, J.; Teixeira, A.; Silva, S.; Oliveira, C.; Dias, M.S. Detecting Nasal Vowels in Speech Interfaces Based on Surface Electromyography. *PLoS ONE* **2015**, *10*, e0127040. [CrossRef]