brain sciences

MDPI

Article

Silent Speech Decoding Using Spectrogram Features Based on Neuromuscular Activities

You Wang [1], Ming Zhang [1], RuMeng Wu [1], Han Gao [1], Meng Yang [2], Zhiyuan Luo [3] and Guang Li [1,*]

[1] State Key Laboratory of Industrial Control Technology, Institute of Cyber Systems and Control, Zhejiang University, Hangzhou 310027, China; king_wy@zju.edu.cn (Y.W.); drystan@zju.edu.cn (M.Z.); 21932107@zju.edu.cn (R.W.); gao_han@zju.edu.cn (H.G.)

[2] Department of Computer Science and Technology, School of Mechanical Electronic and Information Engineering, China University of Mining and Technology, Beijing 100083, China; m.yang@cumtb.edu.cn

[3] Computer Learning Research Centre, Royal Holloway, University of London, Egham Hill, Egham, Surrey TW20 0EX, UK; zhiyuan.luo@cs.rhul.ac.uk

* Correspondence: guangli@zju.edu.cn

Received: 13 May 2020; Accepted: 8 July 2020; Published: 11 July 2020

check for updates

Abstract: Silent speech decoding is a novel application of the Brain–Computer Interface (BCI) based on articulatory neuromuscular activities, reducing difficulties in data acquirement and processing. In this paper, spatial features and decoders that can be used to recognize the neuromuscular signals are investigated. Surface electromyography (sEMG) data are recorded from human subjects in mimed speech situations. Specifically, we propose to utilize transfer learning and deep learning methods by transforming the sEMG data into spectrograms that contain abundant information in time and frequency domains and are regarded as channel-interactive. For transfer learning, a pre-trained model of Xception on the large image dataset is used for feature generation. Three deep learning methods, Multi-Layer Perception, Convolutional Neural Network and bidirectional Long Short-Term Memory, are then trained using the extracted features and evaluated for recognizing the articulatory muscles' movements in our word set. The proposed decoders successfully recognized the silent speech and bidirectional Long Short-Term Memory achieved the best accuracy of 90%, outperforming the other two algorithms. Experimental results demonstrate the validity of spectrogram features and deep learning algorithms.

Keywords: silent speech decoding; neuromuscular signal; spectrogram features; Xception; bidirectional long short-term memory

1. Introduction

Research on Brain–Computer Interfaces (BCI) has a long history [1] and has attracted more attention for its extensive potential in the fields of neural engineering, clinical rehabilitation, daily communication and many other possible applications [2–4]. A typical non-invasive BCI uses electroencephalography (EEG) as it is inexpensive and easy to implement [5]. However, the difficulty in data processing still remains for practical use. One promising approach to address the challenge is the neuromuscular decoding from articulatory muscles [6]. Surface Electromyography (sEMG) captures neuromuscular activities in a non-invasive way like EEG. Besides, it only requires a few channels for signal processing due to the neural pathway from the brain to muscle acting as a primary filter and encoder [7–9].

In the accessible area around the face, surface electrodes are placed on articulatory muscles to obtain speech-related sEMG, both in vocal and silent speech [6,10–12]. Some other techniques are also used in the silent speech recording. Video and ultrasound imaging can record the movements of visible or invisible speech articulators straightforwardly [13,14]. However, they do not work in purely silent speech without any articulator motion.

The primary use of sEMG for silent speech recognition can date back to the mid-1980s, when Sugie in Japan [15] and Morse in the United States [16] demonstrated that sEMG signals contain speech-related information, respectively. Using simple thresholding techniques, Sugie utilized a three-channel electrode to distinguish five Japanese vowels, verifying that they could run in a pilot real-time system [15]. Later, Morse obtained linguistic information from muscle activities of neck and head, successfully distinguishing two words [16]. In the following years, the word number expanded to ten with an accuracy of 70% [17]. However, when it increased to 17, the accuracy dropped to only 35% [18]. In 2001, Chan reported the work of recognizing 10 English numbers based on sEMG during speech, using a wavelet transform feature set with linear discriminant analysis [19]. Later on, a group of researchers utilized sEMG to identify six commands to control an aircraft [20]. Szu-Chen studied continuous audible speech recognition using sEMG, achieving a 32% error rate by decomposing the signal into different feature spaces in time domain [21]. In 2014, Wand used four polar and two bipolar electrodes to capture sEMG to achieve the best average silent speech error rate at 34.7%, where zero-crossing rate, mean value and signal power were extracted [9]. Early in 2018, Kapur reported a wearable silent speech interface to obtain good accuracy around 90%, using a convolutional neural network [6]. Later, Meltzner demonstrated that silent speech was recognized with high accuracy using vocal speech features on a large data set [22].

Although multiple electrodes lead to multi-channel sEMG, previous studies mostly focus on channel-wise features which are extracted on a channel-by-channel basis, while correlations between channels are ignored. Speech is produced by the synergistic work of vocal system and articulatory neuromuscular activities occur along with these physiological processes [23–25]. Even in silent speech, such signals can be recorded and synergistic mechanism exists among the muscles. So, synergistic features from multi-channel sEMG are considered to recognize different words. Xception, originally designed for image classification [26–29], is utilized to process spectrograms of multichannel sEMG to explore the spatial correlation.

In this paper, multi-channel sEMG of silent speech are recorded. Xception is exploited to extract spatial correlative features. Three deep learning methods, Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN) and bidirectional Long Short-Term Memory (bLSTM), are evaluated to decode silent speech.

2. Silent Speech Data

2.1. Capturing Speech-Related sEMG

Studying the relationships between vocalization and articulatory muscles, we select suitable electrode positions around the face [6,8,30–33], as shown in Figure 1. Channels 2 and 5 are bipolar derivation to improve the common-mode rejection ratio (CMRR) while others are derived unipolarity. Channels 1 and 2 record the levator anguli oris while channel 4 captures both the extrinsic tongue and the digastric anterior belly. Channels 3, 5 and 6 record the platysma, the extrinsic tongue and the lateral pterygoid, respectively. Besides, two reference electrodes are placed on the mastoid behind ears.

There is no articulator motion in silent speech, so the amplitude of sEMG is generally below 1 mV, smaller than normal sEMG. The frequency band of silent speech sEMG is always no more than 300 Hz. In our data recording system, the bandwidth is approximately 5 kHz and a 24-bit analog-to-digital converter (ADC) is used. Two resistor–capacitor (RC) filters, including a direct current (DC) filter and a 5 kHz low-pass filter are exploited to eliminate the DC bias and high-frequency interference, respectively. sEMG data are recorded at a sampling rate of 1000 Hz.

Seven students with normal vision and oral expression skills, having no history of mental illness and neurological diseases, 20 to 25 years old (average 22, four males and three females), are recruited as subjects. The experiment named “BCI research based on transmitted neural signals” has been approved by Ethics and Human and Animal Protection Committee of Zhejiang University (Ethical Approval: ZJUEHAPC2019-CSEA01), and strictly follows the Declaration of Helsinki. All collected data are only used for data analysis and the privacy of the participants are firmly protected.

The six-channel sEMG is recorded while the subjects are trained to imagine speaking the labelled words displayed on a computer screen one by one in a defined sequence, which is the meaning of silent speech in this paper. In our experiments, ten Chinese words are selected, including ’噪’, ’1#’, ’2#’, ’前’, ’后’, ’左’, ’右’, ’快’, ’慢’, ’停’ , which mean ’null’, ’No.1’, ’No.2’, ’forward’, ’backward’, ’left’, ’right’, ’accelerate’, ’decelerate’, ’stop’ in English, respectively. In total, 69,296 valid samples for the ten words are recorded, and the label distribution is various, as shown in Table 1. Figure 2 illustrates a valid six-channel sEMG example.

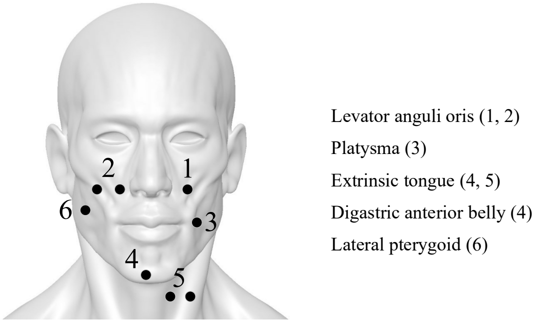


Figure 1. Recording sites around the face and neck. These dedicated positions form an articulator muscular net to decode the silent speech. The sites are cleaned by gel to ensure the impedance is lower than 5 kΩ between electrodes and skin surface.

Table 1. Valid samples.

Label	‘0’	‘1’	‘2’	‘3’	‘4’	‘5’	‘6’	‘7’	‘8’	‘9’
Word	’噪’	’1#’	’2#’	’ 前’	’后’	’左’	’右’	’快’	’慢’	’停’
Samples	7964	6707	6814	6978	6593	6510	6682	6883	7614	6524

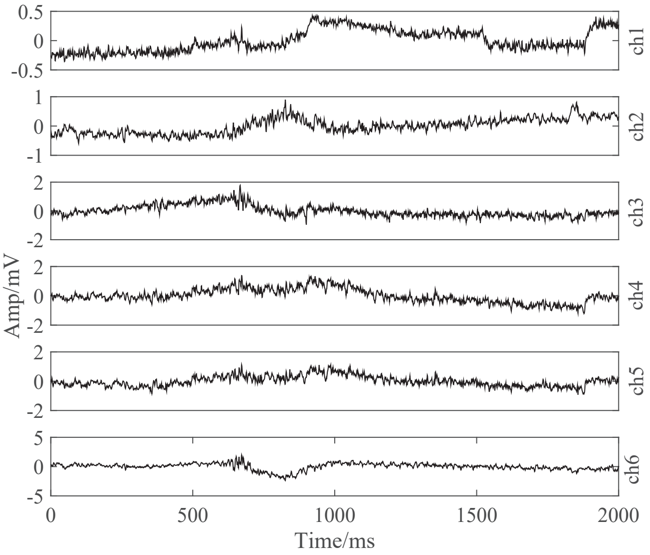


Figure 2. An example of six-channel surface electromyography (sEMG) when imagining to speak ‘decelerate’ in Chinese.

2.2. Preprocessing

An 8th order Butterworth bandpass filter (0.15∼300 Hz) was applied to remove the DC bias and high frequency of sEMG. The power frequency of 50 Hz and its harmonics was filtered by a comb notch filter [6,34–36]. The filtered sEMG is then obtained, as shown in Figure 3b.

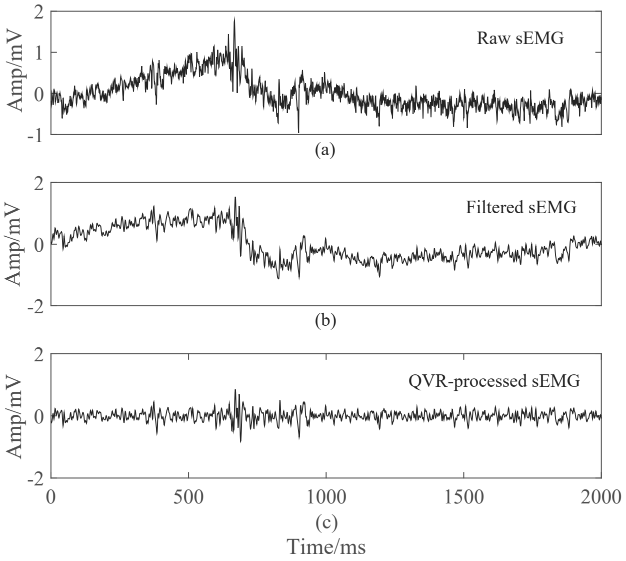


Figure 3. Preprocessing of sEMG. (**a**) An example of raw sEMG, corresponding to channel 2 in Figure 2; (**b**) The sEMG filtered by Butterworth (0.15∼300 Hz) and notch (50 Hz) filters; (**c**) Quadratic Variation Reduction (QVR)-processed sEMG, where the most amplitude change is less than 1 mV.

In order to remove the baseline drift, the Quadratic Variation Reduction (QVR) [37] method is applied:

$$\boldsymbol{z} = [\boldsymbol{I} - (\boldsymbol{I} + \lambda \boldsymbol{D}^T \boldsymbol{D})^{-1}]\tilde{\boldsymbol{z}} \tag{1}$$

where $\tilde{\boldsymbol{z}}$ and $\boldsymbol{z}$ denote the signal before and after using QVR, λ is a constant value ($\lambda = 100$), $\boldsymbol{I}$ represents the identity matrix and $\boldsymbol{D}$ is a $(n-1) \times n$ matrix:

$$\boldsymbol{D} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{bmatrix} \tag{2}$$

where n is the length of $\tilde{\boldsymbol{z}}$.

In Equation (1), $(\boldsymbol{I} + \lambda \boldsymbol{D}^T \boldsymbol{D})$ is a symmetric, positive-definite, tridiagonal matrix, which can be solved efficiently. The effect is shown in Figure 3c where it can be seen that most wander part is removed.

3. Processing Methods

In order to extract time–frequency features effectively, the original six-channel sEMG in the time domain were transformed into the frequency domain, creating a spectrogram which is represented as an image. The state-of-the-art model Xception was selected for extracting image features, which were then decoded by MLP, CNN and bLSTM, respectively. Figure 4 describes the processes to decode sEMG.

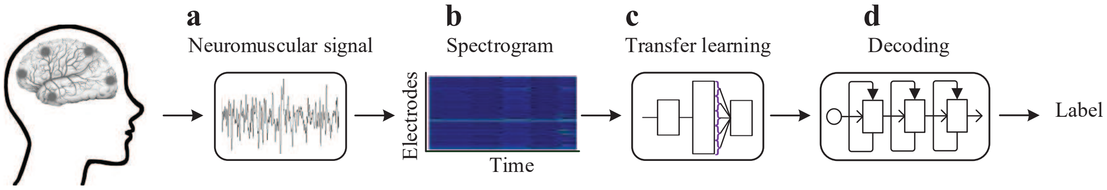


Figure 4. Silent speech decoding. (**a**): The neuromuscular activities are captured by surface electrodes; (**b**): All data are transformed into spectrograms by short-time Fourier transform (STFT); (**c**): Transfer learning method is used to extract features from spectrograms; (**d**): Neural networks decode multi-channel sEMG using the extracted features.

3.1. Spectrogram Images

The spectrogram of a signal sequence is the visual representation of the magnitude of the time-dependent Fourier Transform (FT) versus time, also known as the short-time Fourier transform (STFT) [38–40]. It describes the spectral details in time-frequency domain.

$$\boldsymbol{Spectrogram} = |STFT(x)|^2. \tag{3}$$

The spectrogram was calculated by Equation (3) [38], where the parameters of [window, window length, sample rate, overlap, FFT length] were specified as [hanning, 512, 1000 Hz, 50%, 64]. An example of a spectrogram image is shown in Figure 5. The images associate with each other, reflecting sEMG spatial relationships in the frequency domain. Inspired by short video streams, the images were treated as a fixed-size video. Then, the silent speech decoding becomes a video classification, explored by deep learning methods.

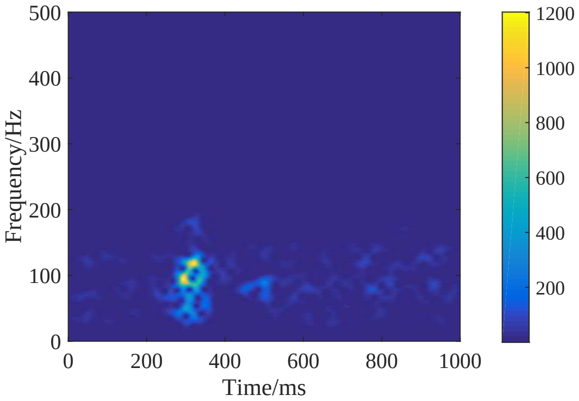


Figure 5. An example of a spectrogram image.

3.2. Feature Extraction

To explore sEMG spatial features, transfer learning with Xception is used. It is a deep learning image classifier using depthwise separable convolution layers with residual connections, which has been pre-trained on large scale images [26]. After input, data using only pointwise convolution (1 × 1 convolution) create separate convolution sizes of 3 × 3 without average pooling, which proceeds in nonoverlapping sections of the output channels to then be fed-forward for concatenation [26,27]. The model demonstrates a strong ability to generalize to images outside the original dataset via transfer learning, such as feature extraction and fine-tuning. Fine-turning is done by training all weights with a smaller learning rate, removing and updating some biased weights from the original network.

The spectrogram images have various shapes and are scaled to 299×299. Xception model outputs 1000 features for each image, therefore 1000 × 6 = 6000 features are obtained for one sEMG sample. All samples are processed using Xception to generate a large feature set.

3.3. Decoder Design

Three deep learning methods, namely MLP, CNN and bLSTM, are explored using the above feature set. Their structure and parameter details are designed in this section [41–43]. Figure 6 illustrates our decoding process, where parts (c)∼(g) represent the common structures and components for the three models, except that different hidden layers and parameters are used in each model.

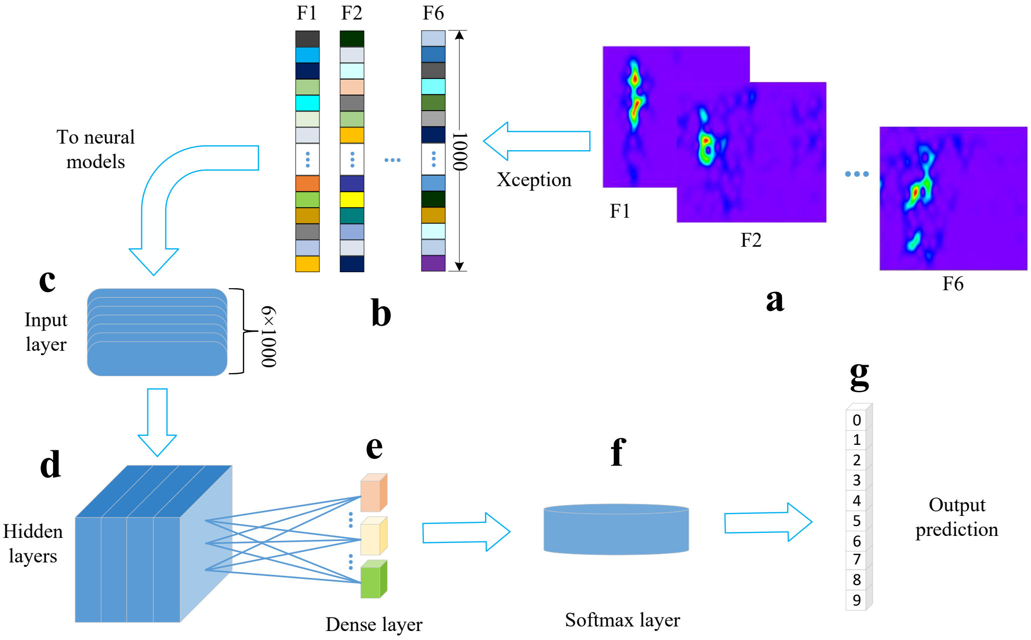


Figure 6. Decoding processes. (**a**) Spectrogram images. (**b**) Feature set extracted by Xception. (**c**) Input layer of neural networks. (**d**) Hidden layers of neural networks. (**e**) Fully connected dense layer. (**f**) Softmax layer as the output layer. (**g**) The predicted labels we obtain from the models.

3.3.1. MLP

Multi-Layer Perceptron (MLP) is a common Artificial Neural Network (ANN). In addition to the input and output layers, there can be multiple hidden layers. MLP can also be thought of as a directed graph consisting of multiple layers, each fully connected to the next layer [44,45].

Figure 7 illustrates the MLP structure where the 'dense' layer connects each input unit with each output unit of the layer to learn and update the weights. 'Dropout' regularization is used to help prevent overfitting as it randomly drops out input units with a fixed rate during parameter tuning [46]. 'Softmax' calculates predicted label probabilities at the output layer and then outputs the label with the maximum probability. The loss function defined in this method is cross-entropy loss.

Each hidden layer uses a non-linear activation function to enhance the performance of the neural network and solve the linear inseparable problem. Commonly used activation functions are sigmoid, tanh, and rectified linear unit (ReLU). ReLU is used in the MLP as the derivative of ReLU is always 1 in the positive interval, alleviating the gradient disappearance and gradient explosion problems. In addition, ReLU has a much faster convergence than sigmoid and tanh.

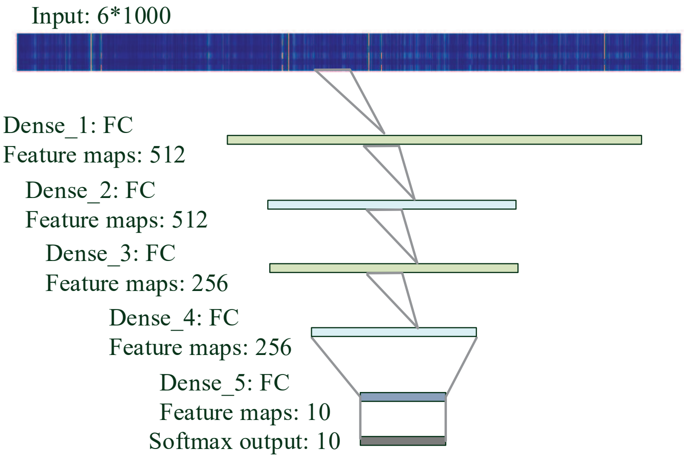


Figure 7. MultiLayer Perceptron (MLP) architecture to decode silent speech. A feature vector goes through the layers and a digital (from 0 to 9) will be output.

3.3.2. CNN

Convolutional Neural Network (CNN) features with local connections and shared weights, making it very popular and successful in image classification problems. The core operation of CNN is mathematical convolution which consists of filters. The convolution is applied on the input data to produce a feature map. Specifically designed filters can extract features via convolution [47–49].

The CNN structure is shown in Figure 8, where two convolutional layers (Conv1 and Conv2) with different filters are used to create specific feature maps. The pooling layer provides downsampling to reduce the size of features and also helps prevent overfitting. Max pooling that calculates the maximum value for each patch is used in our CNN architecture.

In the neural networks, the output of the first layer feeds into the second layer, and the output of the second layer feeds into the third, and so on. When the parameters of a layer change, so does the distribution of inputs to subsequent layers [50], which is described as an internal covariate shift. These shifts in input distribution can be problematic for neural networks, especially deep neural networks that have a large number of layers [51]. Batch normalization, a technique to standardize the inputs to a layer and reduce unwanted shifts to speed up training [52], is used in the CNN model.

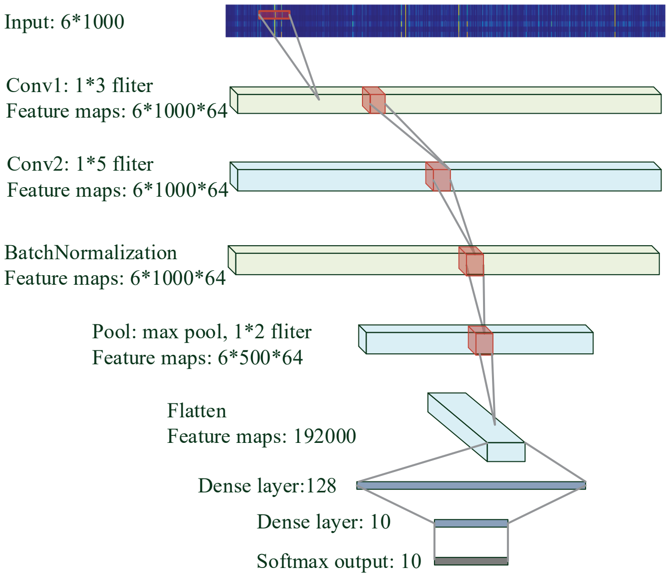


Figure 8. Convolutional Neural Network (CNN) architecture to decode silent speech.

3.3.3. bLSTM

Recurrent Neural Network (RNN) is well-known for processing sequence data, and has made many significant accomplishments in natural language processing applications. Unlike MLP and CNN, the output of each hidden layer in RNN are stored as memory and can be considered as another input, by which it allows information to persist [47,49,53]. However, RNNs suffer from short-term memory. If a sequence is long enough, they will have a hard time carrying information from earlier time steps to later ones. During back propagation, the gradient vanishing in RNN is a serious problem when learning long-term dependencies [53]. The gradient shrinks in back propagation and it does not contribute much to learning if it becomes extremely small [54].

LSTM, a special kind of RNN, addresses this issue by considering that both memory and input operations are addition only. As a result, it is capable of learning long-term dependencies [49]. The core concept of LSTM is the cell state and its various gates. The cell state acts as a transport highway that transfers relative information all the way down the sequence chain. LSTM has the ability to remove or add information by gates. There are the forget gate, input gate and output gate to regulate the flow of information inside the LSTM unit and learn which data in a sequence is important to keep or dismiss.

bLSTM, including forward LSTM and backward LSTM, captures bidirectional semantic dependencies [44,54]. For six-channel sEMG, bLSTM tends to be a suitable classifier as it can effectively model bidirectional dependencies. Figure 9 shows details of the bLSTM architecture, consisting of three bidirectional layers, two dense layers and one softmax output layer.

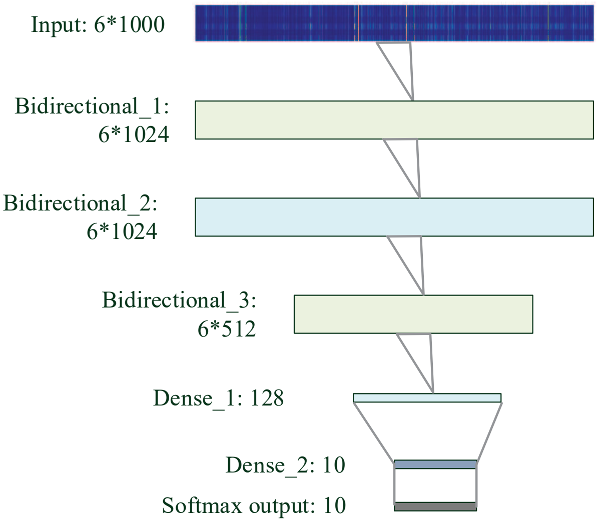
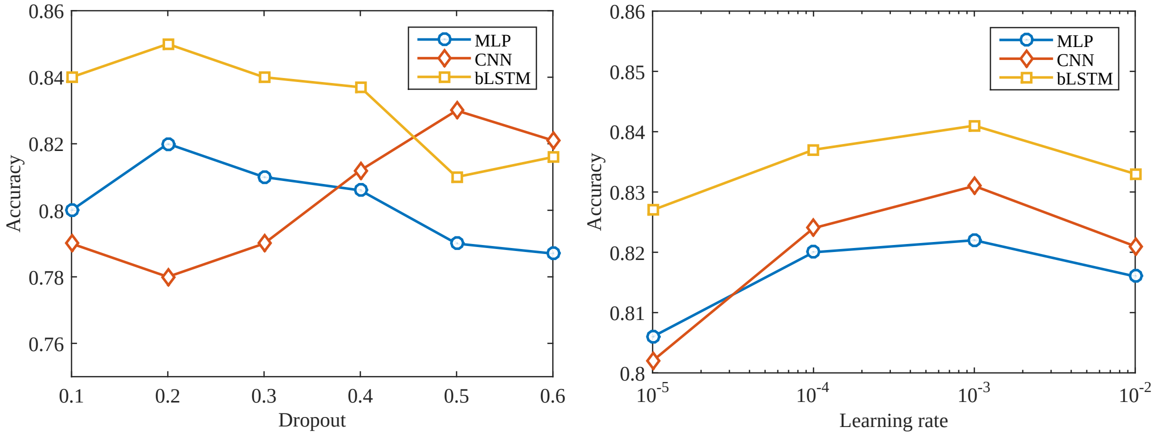


Figure 9. Bidirectional Long Short-Term Memory (bLSTM) architecture to decode silent speech.

4. Results

4.1. Decoder Optimization

For model training and testing purposes, the data set is randomly split at the ratio of training: validate: test = 7:2:1. The structures and parameters of MLP, CNN and bLSTM are optimized based on a series of trials. There are a number of experiments that have been implemented to explore optimal hyperparameters, including dropout rate, learning rate, and network depth. Figure 10a presents that the best dropout rates for MLP and bLSTM are 0.2 while for CNN is 0.5. Learning rate controls how much the weights in neural networks are adjusted with respect to the loss gradient [55,56]. To explore a better initial learning rate in the decaying scheme, experiments are implemented. Figure 10b indicates the initial learning rate of 1×10^{-3} suits for all the three methods.



(**a**) (**b**)

Figure 10. Dropout and learning rate optimization.

Different depths of these networks are also tried while other parameters remain the same. More layers lead to a decrease of prediction or overfitting whereas fewer layers may not be sufficiently trained. Figures 7–9 provide more details of the final topologies with the suitable depth.

4.2. Decoding Results

The features are trained, validated and tested by MLP, CNN and bLSTM, respectively. The key hyperparameters of the deep learning models are displayed in Table 2. The same initial learning rate, activation function and batch size are used in the three decoders, while different optimizers and dropout rates are applied.

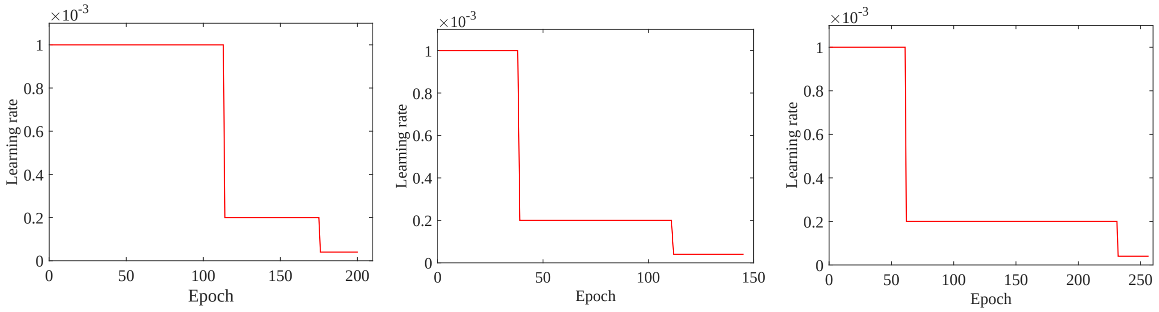
Table 2. Hyperparameters.

Model	Optimizer	Dropout	Learning Rate	Activation	Batch Size
MLP	adam	0.2	1×10^{-3}	ReLU	32
CNN	adadelta	0.5	1×10^{-3}	ReLU	32
bLSTM	rmsprop	0.2	1×10^{-3}	ReLU	32

MLP, CNN and bLSTM are implemented in Keras (on top of TensorFlow), which offers many flexible functional APIs to build and optimize deep learning structures and parameters. Batch normalization is applied for all models to obtain smaller training and validation loss. In particular, the function 'ReduceLROnPlateau' is called to reduce learning rate, with factor = 0.2, patience = 20 and min_lr = 0.5×10^{-6}. In early stopping, patience is set to 80, which means training is stopped if the loss does not decrease after 80 epochs.

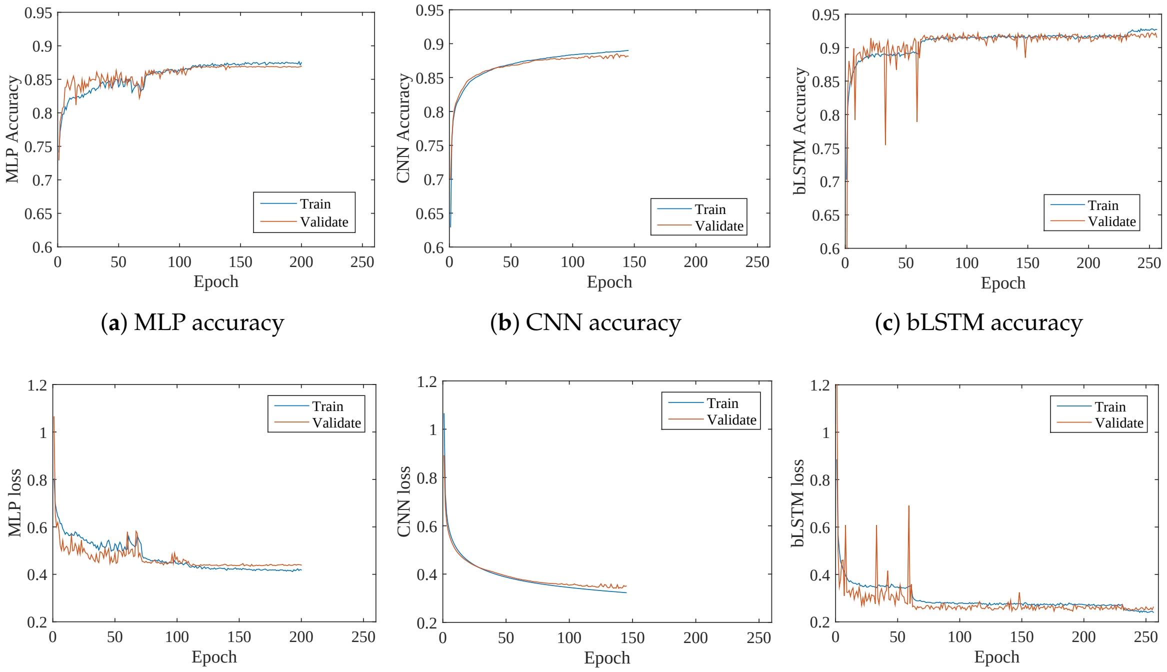
Figure 11 shows the learning rate changes along with epochs during the training. All the three models are initialized with the same learning rate which is then decayed in different epochs. bLSTM takes more than 250 epochs in model training, while that of MLP is smaller and CNN consumes the least number of epochs.

Training profiles are provided in Figure 12. Figure 12a,d give the training details of MLP, where the accuracy becomes stable around 150 epochs and the validation loss stays about 0.45. In Figure 12b,e, CNN training achieves a little better validation results than MLP but a large number of epochs is required. bLSTM shows the best validation accuracy of 0.92 in Figure 12c and the lowest validation loss of 0.26 in Figure 12f, however, its computational efficiency is not as good as those of MLP and CNN since bLSTM needs a large number of epochs to complete the training. The validation performance lines generally follow the training processes, which means the models are generally well-trained without obvious overfitting or underfitting.



(**a**) MLP (**b**) CNN (**c**) bLSTM

Figure 11. Learning rate of prediction in decoders.



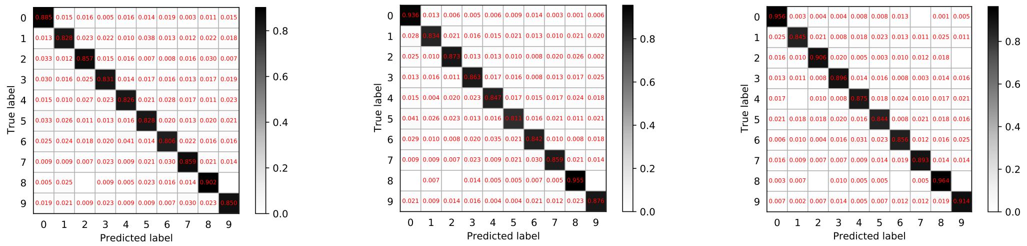
(**a**) MLP accuracy (**b**) CNN accuracy (**c**) bLSTM accuracy

(**d**) MLP loss (**e**) CNN loss (**f**) bLSTM loss

Figure 12. Training profile on the feature set by three deep learning models. (**a**) and (**d**): training on MLP. (**b**) and (**e**): training on CNN. (**c**) and (**f**): training on bLSTM. Both training and validation results are shown in the above sub-figures.

The accuracy of MLP, CNN and bLSTM on the test set is 0.85, 0.87 and 0.90, respectively. Both training and test results indicate that bLSTM achieves the best performance among the three methods, though it takes a longer time to train.

The confusion matrix is computed to show more prediction details on the test set, as is shown in Figure 13. Labels 0 and 8 achieve the highest accuracy in all test predictions while labels 1, 5 and 6 have relatively low accuracy. Except for label 5, the accuracy of all others increases from Figure 13a,c. Samples are more likely to be classified as labels 0 or 8. In addition, all three decoders have an equal difficulty in distinguishing label 4 and label 6. This may be caused by similar neuromuscular activities.



(**a**) MLP (**b**) CNN (**c**) bLSTM

Figure 13. Confusion matrices of the three decoders.

5. Discussion

The valid samples for each label are various, due to (1) the impedance between an electrode and skin surface always changes in different experiments, even for the same participant; (2) the inherent differences in the speech intention of participants; and (3) the different responses from the neuromuscular activities to individual words in silent speech recording. The data set, as shown in Table 1, is still acceptable though small label imbalance exists, because each label is fully trained. The impedance reduction and preprocessing algorithm optimization will be further studied to increase the rate of valid samples.

MLP, CNN and bLSTM are trained and applied to decode the sEMG on the same platform (Intel i5-7400 CPU @ 3 GHz). bLSTM obtains highest accuracy around 0.92, with largest time consumption of almost 10 h. For CNN, the performance is not as good as bLSTM (0.88), but it consumes the least time (6 h) for proper model training. Though MLP takes less time (8 h) than bLSTM, its accuracy of 0.87 is worst among the three models. The bi-directional structure in bLSTM can generate better decoding results than MLP and CNN. Therefore, bLSTM suits the silent speech recognition if time consumption is less important. In test experiments, it takes no more than 50 ms to predict a new sEMG sample for the three models, which means instant prediction can be obtained to satisfy a real-time system.

The technology of silent speech decoding can be output in two forms, text code and synthetic speech [9,57]. It is up to the practical requirements. The speech pattern only appears in sEMG form regardless of audible or silent speech, so the privacy is ensured by the subject.

Silent speech decoding investigated in this paper is promising in possible applications: medical prostheses that help people with speech disabilities; hands-free peripheral device control; communication in privacy or noisy ambience [22,57–59]. The accuracy of single word is not high enough for piratical use. Communication also requires more complicated expression than the single words. Semantic dependency may help silent speech recognition in such potential applications, so phrases or even sentences may need to be researched.

Currently, 10 electrodes (2 for ground, 2 pairs of bipolar and 4 monopolar electrodes) for 6 channels are needed, and an integrated electrode array will be developed to improve the wearability. Furthermore, the electrode positions and channel number might be optimized to improve the performance and simplify the data collecting device. Online learning is another possible future research, as it is useful in data augmentation.

6. Conclusions

In this paper, it is demonstrated that spectrogram features combined with deep learning models can be applied to the silent speech decoding task, where bLSTM outperforms other methods. Result analysis indicates that synergic information hidden in multi-channel sEMG can provide useful features for recognition. It is suggested that the synergic exploration in silent speech decoding should be extended to phrases or even sentences, not only limited to a single word.

Author Contributions: conceptualization, Y.W. and G.L.; methodology, M.Z. and Z.L.; software, Y.W. and Z.M.; validation, M.Z.; formal analysis, Y.W. and M.Z.; investigation, M.Y. and M.Z.; resources, Z.L.; data curation, M.Z., R.W. and H.G.; writing—original draft preparation, M.Z., R.W. and H.G.; writing—review and editing, G.L. and M.Z.; visualization, G.L.; supervision, Y.W.; project administration: G.L.; funding acquisition, G.L. and Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Zhejiang University Education Foundation Global Partnership Fund and the Fundamental Research Funds for the Central Universities.

Acknowledgments: The authors would like to thank all participants who contributed to our project, and special thanks go to Wei Zhang, Huiyan Li, and Qing Ai for their kindly help for data collection.

Conflicts of Interest: The authors declare no conflict of interest and do not hold relevant patents.

References

1. Vidal, J.J. Toward direct brain-computer communication. *Annu. Rev. Biophys. Bioeng.* **1973**, *2*, 157–180. [CrossRef] [PubMed]
2. Pfurtscheller, G.; Flotzinger, D.; Kalcher, J. Brain-computer interface-a new communication device for handicapped persons. *J. Microcomp. Appl.* **1993**, *16*, 293–299. [CrossRef]
3. Ang, K.K.; Chua, K.S.G.; Phua, K.S.; Wang, C.; Chin, Z.Y.; Kuah, C.W.K.; Low, W.; Guan, C. A randomized controlled trial of EEG-based motor imagery brain-computer interface robotic rehabilitation for stroke. *Clin. EEG Neurosci.* **2015**, *46*, 310–320. [CrossRef] [PubMed]
4. Mahmood, M.; Mzurikwao, D.; Kim, Y.S.; Lee, Y.; Mishra, S.; Herbert, R.; Duarte, A.; Ang, C.S.; Yeo, W.H. Fully portable and wireless universal brain–machine interfaces enabled by flexible scalp electronics and deep learning algorithm. *Nat. Mach. Intell.* **2019**, *1*, 412–422. [CrossRef]
5. Ramadan, R.A.; Vasilakos, A.V. Brain computer interface: Control signals review. *Neurocomputing* **2017**, *223*, 26–44. [CrossRef]
6. Kapur, A.; Kapur, S.; Maes, P. Alterego: A personalized wearable silent speech interface. In *23rd International Conference on Intelligent User Interfaces*; ACM: New York, NY, USA, 2018; pp. 43–53.
7. Yau, W.C.; Arjunan, S.P.; Kumar, D.K. Classification of voiceless speech using facial muscle activity and vision based techniques. In Proceedings of the TENCON 2008-2008 IEEE Region 10 Conference, Hyderabad, India, 19–21 November 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 1–6.
8. Schultz, T.; Wand, M. Modeling coarticulation in EMG-based continuous speech recognition. *Speech Commun.* **2010**, *52*, 341–353. [CrossRef]
9. Wand, M.; Janke, M.; Schultz, T. Tackling speaking mode varieties in EMG-based speech recognition. *IEEE Trans. Biomed. Eng.* **2014**, *61*, 2515–2526. [CrossRef]
10. Wand, M.; Schultz, T. Speaker-adaptive speech recognition based on surface electromyography. In *International Joint Conference on Biomedical Engineering Systems and Technologies*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 271–285.
11. Deng, Y.; Colby, G.; Heaton, J.T.; Meltzner, G.S. Signal processing advances for the MUTE sEMG-based silent speech recognition system. In Proceedings of the MILCOM 2012-2012 IEEE Military Communications Conference, Orlando, FL, USA, 29 October–1 November 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 1–6.
12. Soon, M.W.; Anuar, M.I.H.; Abidin, M.H.Z.; Azaman, A.S.; Noor, N.M. Speech recognition using facial sEMG. In Proceedings of the 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Kuching, Malaysia, 12–14 September 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–5.
13. Denby, B.; Schultz, T.; Honda, K.; Hueber, T.; Gilbert, J.M.; Brumberg, J.S. Silent speech interfaces. *Speech Commun.* **2010**, *52*, 270–287. [CrossRef]
14. Hofe, R.; Ell, S.R.; Fagan, M.J.; Gilbert, J.M.; Green, P.D.; Moore, R.K.; Rybchenko, S.I. Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing. *Speech Commun.* **2013**, *55*, 22–32. [CrossRef]
15. Sugie, N.; Tsunoda, K. A speech prosthesis employing a speech synthesizer-vowel discrimination from perioral muscle activities and vowel production. *IEEE Trans. Biomed. Eng.* **1985**, *BME-32*, 485–490. [CrossRef] [PubMed]
16. Morse, M.S.; O'Brien, E.M. Research summary of a scheme to ascertain the availability of speech information in the myoelectric signals of neck and head muscles using surface electrodes. *Comput. Biol. Med.* **1986**, *16*, 399–410. [CrossRef]

17. Morse, M.S.; Day, S.H.; Trull, B.; Morse, H. Use of myoelectric signals to recognize speech. In *Images of the Twenty-First Century. Proceedings of the Annual International Engineering in Medicine and Biology Society, Seattle, WA, USA, 9–12 November 1989*; IEEE: Piscataway, NJ, USA, 1989; pp. 1793–1794.
18. Morse, M.; Gopalan, Y.; Wright, M. Speech recognition using myoelectric signals with neural networks. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Orlando, FL, USA 31 October–3 November 1991*; IEEE: Piscataway, NJ, USA, 1991; Volume 13, pp. 1877–1878.
19. Chan, A.D.; Englehart, K.; Hudgins, B.; Lovely, D.F. Myo-electric signals to augment speech recognition. *Med. Biol. Eng. Comput.* **2001**, *39*, 500–504. [CrossRef] [PubMed]
20. Jorgensen, C.; Lee, D.D.; Agabont, S. Sub auditory speech recognition based on EMG signals. In Proceedings of the International Joint Conference on Neural Networks, 2003, Portland, OR, USA, 20–24 July 2003; IEEE: Piscataway, NJ, USA, 2003; Volume 4, pp. 3128–3133.
21. 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; pp. 573–576.
22. Meltzner, G.S.; Heaton, J.T.; Deng, Y.; De Luca, G.; Roy, S.H.; Kline, J.C. Development of sEMG sensors and algorithms for silent speech recognition. *J. Neural Eng.* **2018**, *15*, 046031. [CrossRef] [PubMed]
23. Martini, F.; Nath, J.L.; Bartholomew, E.F.; Ober, W.C.; Ober, C.E.; Welch, K.; Hutchings, R.T. *Fundamentals of Anatomy & Physiology*; Pearson Benjamin Cummings: San Francisco, CA, USA, 2006; Volume 7.
24. Marieb, E.N.; Hoehn, K. *Human Anatomy & Physiology*, 9th ed.; Pearson: London, UK, 2013; pp. 276–482.
25. Schultz, T.; Wand, M.; Hueber, T.; Krusienski, D.J.; Herff, C.; Brumberg, J.S. Biosignal-Based Spoken Communication: A Survey. *IEEE Trans. Audio. Speech. Lang. Process.* **2017**, *25*, 2257–2271. [CrossRef]
26. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; IEEE: NewYork, NY, USA, 2017; pp. 1251–1258.
27. Jinsakul, N.; Tsai, C.F.; Tsai, C.E.; Wu, P. Enhancement of Deep Learning in Image Classification Performance Using Xception with the Swish Activation Function for Colorectal Polyp Preliminary Screening. *Mathematics* **2019**, *7*, 1170. [CrossRef]
28. Yang, L.; Chen, X.; Tao, L. Acoustic scene classification using multi-scale features. In Proceedings of the Workshop on DCASE 2018, Surrey, UK, 19–20 November 2018.
29. Yang, L.; Yang, P.; Ni, R.; Zhao, Y. Xception-Based General Forensic Method on Small-Size Images. In *Advances in Intelligent Information Hiding and Multimedia Signal Processing*; Springer: Berlin, Germany, 2020; pp. 361–369.
30. Hermens, H.J.; Freriks, B.; Disselhorstklug, C.; Rau, G. Development of recommendations for SEMG sensors and sensor placement procedures. *J. Electromyogr. Kinesiol.* **2000**, *10*, 361–374. [CrossRef]
31. Roberts, A. *Human Anatomy: The Definitive Visual Guide*; Dorling Kindersley Ltd.: London, UK, 2016; pp. 50–65.
32. Kenneth, S.S. *Anatomy & Physiology: The Unity of Form and Function*; McGraw-Hill: New York, NY, USA, 2017; pp. 307–570.
33. Zhang, M.; Wang, Y.; Wei, Z.; Yang, M.; Luo, Z.; Li, G. Inductive conformal prediction for silent speech recognition. *J. Neural Eng.* **2020**, in press. [CrossRef]
34. Maier-Hein, L.; Metze, F.; Schultz, T.; Waibel, A. Session independent non-audible speech recognition using surface electromyography. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, 2005, San Juan, Puerto Rico, 27 November–1 December 2005; IEEE: Piscataway, NJ, USA, 2005; pp. 331–336.
35. Stepp, C.E.; Heaton, J.T.; Rolland, R.G.; Hillman, R.E. Neck and face surface electromyography for prosthetic voice control after total laryngectomy. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2009**, *17*, 146–155. [CrossRef]
36. Hakonen, M.; Piitulainen, H.; Visala, A. Current state of digital signal processing in myoelectric interfaces and related applications. *Biomed. Signal Process. Control* **2015**, *18*, 334–359. [CrossRef]
37. Fasano, A.; Villani, V. Baseline wander removal for bioelectrical signals by quadratic variation reduction. *Signal Process.* **2014**, *99*, 48–57. [CrossRef]
38. Sairamya, N.; Susmitha, L.; George, S.T.; Subathra, M. Hybrid Approach for Classification of Electroencephalographic Signals Using Time–Frequency Images With Wavelets and Texture Features. In *Intelligent Data Analysis for Biomedical Applications*; Elsevier: Amsterdam,The Netherlands, 2019; pp. 253–273.
39. Huang, J.; Chen, B.; Yao, B.; He, W. ECG Arrhythmia Classification Using STFT-Based Spectrogram and Convolutional Neural Network. *IEEE Access* **2019**, *7*, 92871–92880. [CrossRef]

40. Pandey, A.; Wang, D. Exploring Deep Complex Networks for Complex Spectrogram Enhancement. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 6885–6889.
41. Géron, A. *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2017.
42. Xianshun, C. Keras Implementation of Video Classifier. Available online: https://github.com/chen0040/keras-video-classifier (accessed on 13 April 2018).
43. Anumanchipalli, G.K.; Chartier, J.; Chang, E.F. Speech synthesis from neural decoding of spoken sentences. *Nature* **2019**, *568*, 493. [CrossRef] [PubMed]
44. Orhan, U.; Hekim, M.; Ozer, M. EEG signals classification using the K-means clustering and a multilayer perceptron neural network model. *Expert Syst. Appl.* **2011**, *38*, 13475–13481. [CrossRef]
45. Tang, J.; Deng, C.; Huang, G.B. Extreme learning machine for multilayer perceptron. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *27*, 809–821. [CrossRef]
46. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
47. Shin, H.C.; Roth, H.R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; Summers, R.M. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imag.* **2016**, *35*, 1285–1298. [CrossRef]
48. Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; Xu, W. Cnn-rnn: A unified framework for multi-label image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 2285–2294.
49. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep learning*; MIT Press: Cambridge, MA, USA, 2016.
50. Bjorck, N.; Gomes, C.P.; Selman, B.; Weinberger, K.Q. Understanding batch normalization. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; Curran Associates, Inc.: Red Hook, NY, USA, 2018; pp. 7694–7705.
51. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167 .
52. Santurkar, S.; Tsipras, D.; Ilyas, A.; Madry, A. How does batch normalization help optimization? In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; Curran Associates, Inc.: Red Hook, NY, USA, 2018; pp. 2483–2493.
53. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
54. Sak, H.; Senior, A.; Beaufays, F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv* **2014**, arXiv:1402.1128 .
55. Zeiler, M.D. Adadelta: an adaptive learning rate method. *arXiv* **2012**, arXiv:1212.5701 .
56. Yu, C.; Qi, X.; Ma, H.; He, X.; Wang, C.; Zhao, Y. LLR: Learning learning rates by LSTM for training neural networks. *Neurocomputing* **2020**, *394*, 41–50. [CrossRef]
57. Janke, M.; Diener, L. Emg-to-speech: Direct generation of speech from facial electromyographic signals. *IEEE Trans. Audio. Speech. Lang. Process.* **2017**, *25*, 2375–2385. [CrossRef]
58. Denby, B.; Chen, S.; Zheng, Y.; Xu, K.; Yin, Y.; Leboullenger, C.; Roussel, P. Recent results in silent speech interfaces. *J. Acoust. Soc. Am.* **2017**, *141*, 3646. [CrossRef]
59. Cler, M.J.; Nieto-Castanon, A.; Guenther, F.H.; Stepp, C.E. Surface electromyographic control of speech synthesis. In Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, USA, 26–30 August 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 5848–5851.