Article

A Speech Command Control-Based Recognition System for Dysarthric Patients Based on Deep Learning Technology

Yu-Yi Lin ¹, Wei-Zhong Zheng ¹, Wei Chung Chu ¹, Ji-Yan Han ¹, Ying-Hsiu Hung ¹, Guan-Min Ho ², Chia-Yuan Chang ² and Ying-Hui Lai ^{1,*}

- ¹ Department of Biomedical Engineering, National Yang Ming Chiao Tung University, No. 155, Sec. 2, Taipei 112, Taiwan; lin3000y@gmail.com (Y.-Y.L.); s1010654@gm.ym.edu.tw (W.-Z.Z.); abc55169@gmail.com (W.C.C.); jjeerry2006com@gmail.com (J.-Y.H.); wer850718906@gmail.com (Y.-H.H.)
- ² A Prevent Medical Inc., 7F, No.520, 5 Sec, ZhongShan N. Rd., Shilin Dist., Taipei 11141, Taiwan; mdhgm@apreventmed.com (G.-M.H.); mark@apreventmed.com (C.-Y.C.)
- * Correspondence: yh.lai@gm.ym.edu.tw; Tel.: +886-2-28267021

Abstract: Voice control is an important way of controlling mobile devices; however, using it remains a challenge for dysarthric patients. Currently, there are many approaches, such as automatic speech recognition (ASR) systems, being used to help dysarthric patients control mobile devices. However, the large computation power requirement for the ASR system increases implementation costs. To alleviate this problem, this study proposed a convolution neural network (CNN) with a phonetic posteriorgram (PPG) speech feature system to recognize speech commands, called CNN-PPG; meanwhile, the CNN model with Mel-frequency cepstral coefficient (CNN-MFCC model) and ASR-based systems were used for comparison. The experiment results show that the CNN-PPG system provided 93.49% accuracy, better than the CNN-MFCC (65.67%) and ASR-based systems (89.59%). Additionally, the CNN-PPG used a smaller model size comprising only 54% parameter numbers compared with the ASR-based system; hence, the proposed system could reduce implementation costs for users. These findings suggest that the CNN-PPG system could augment a communication device to help dysarthric patients control the mobile device via speech commands in the future.

Keywords: mobile device; dysarthric speech; speech command recognition (SCR); signal processing; health care; Internet of Things (IoT); deep learning

Citation: Lin, Y.-Y.; Zheng, W.-Z.; Chu, W.C.; Han, J.-Y.; Hung, Y.-H.; Ho, G.-M.; Chang, C.-Y.; Lai, Y.-H. A Speech Command Control-Based Recognition System for Dysarthric Patients Based on Deep Learning Technology. *Appl. Sci.* **2021**, *11*, 2477. <https://doi.org/10.3390/app11062477>

Academic Editors: Arjan Kuijper and Marcos Hortal

Received: 10 December 2020

Accepted: 4 March 2021

Published: 10 March 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Dysarthric speaking is often associated with aging as well as with medical conditions, such as cerebral palsy (CP) and amyotrophic lateral sclerosis (ALS) [1]. It is a motor speech disorder caused by muscle weakness or lack of control and often makes someone's speech unclear; hence, patients cannot communicate well with people (or machines). Currently, augmentative and alternative communication (AAC) systems are used to improve patient communication capabilities, such as communication boards [2], head tracking [3], gesture controlling [4], and eye-tracking [5] technologies. Previous studies have shown that these systems provided benefits to help patients communicate with people; however, there is still room for improvement. For example, communication using these devices is often slow and unnatural for dysarthric patients [6]; hence, it affects the communication performance of dysarthric patients directly. To overcome these issues, many studies [7] have proposed speech command recognition (SCR) systems that can help patients control devices via their voice, such as automatic speech recognition (ASR) systems [8] and acoustic pattern recognition technologies [9].

For SCR systems used in dysarthric patients, one challenge is the phonetic variation of speech [1,10–12]. Phonetic variation in dysarthric patients is a common issue, caused by limitations from neurological injury to the motor component of the motor-speech system [13]. To alleviate phonetic variation of a dysarthric patient's speech, most related

works on the recognition of dysarthric speech have been focused on acoustic modeling to obtain suitable acoustic cues for dysarthric speech. Hasegawa-Johnson et al. [14] evaluate recognition performance for dysarthric speech compared with automatic speech recognition (ASR) systems based on Gaussian mixture model–hidden Markov models (GMM–HMMs) and support vector machines (SVMs) [15]. The experimental results showed that the HMM-based model may provide robustness against large-scale word-length variances; meanwhile, the SVM-based model can alleviate the effect of deletion or reduction in consonants. Rudzicz et al. [16,17] investigated acoustic models of GMM–HMM, conditional random field, SVM, and artificial neural networks (ANNs) [17], and the results showed that the ANNs provided higher accuracy than other models.

Recently, deep learning technology has been widely used in many fields [18–20] and has proven it can provide better performance than conventional classification models in SCR tasks. For example, Snyder et al. [21] apply a deep neural network (DNN) with a data augmentation technique to perform ASR. They achieve better performance in the x-vector system; however, it is not helpful as an i-vector extractor. Fathima et al. [22] applied a multilingual Time Delay Neural Network (TDNN) system that combined acoustic modeling and language specific information to increase ASR performance. The experimental results showed that the TDNN-based ASR system achieved suitable performance, as the word error rate was 16.07% in this study.

Although the ASR-based approach is a classical technology [8] in the dysarthric SCR task, other studies indicate that ASR systems still need huge improvements for severely dysarthric patients (e.g., cerebral palsy or stroke) [23–25]. This may be because ASR systems are trained without including dysarthric speech [14,23,26,27]. Therefore, studies have tried to modify the approach of the ASR system to achieve higher performance. For example, Hawley et al. [28] suggested that a small-vocabulary speaker-dependent recognition system (i.e., the personalized SCR system in this study, which the dysarthric patients need to record their speech) can be more effective for severely dysarthric users in SCR tasks. Farooq et al. [29] applied the wavelet technique [30] to transform the acoustic data to achieve speech recognition. In the experiment, they found that the wavelet technique can achieve better performance than the traditional transform technique Mel-frequency cepstral coefficient (MFCC) in a voiced stops situation; however, MFCC showed better performance in other situations, such as with voiced fricatives. Shahamiri et al. [8] used the best-performing set MFCC [31,32] with artificial neural networks to perform speaker-independent ASR. The experiment results showed an average 68.4% word recognition rate with the dysarthric speaker-independent ASR model and 95.0% word recognition rate in speaker-dependent ASR systems. Park et al. [21] used the data augmentation approach called SpecAugment to improve the ASR performance, and the results showed that this approach could achieve 6.8% word error rate (WER). Yang et al. [25] applied cycle-consistent adversarial training methods to improve dysarthric speech. They achieved a lower WER (33.4%) compared with automatic speech recognition of the generated utterance on a held-out test set. Recognition performance, such as with the above systems, allows users to individually train a system using their speech, thus making it possible to account for the variations in dysarthric speech [28]. Although these approaches can provide suitable performance in this task, there are some issues to overcome, including privacy (i.e., the recorded data uploaded to the server in most ASR systems) and higher computing power needed to use the ASR system. Thus, edge computing-based SCR systems, such as those using acoustic pattern recognition technologies [33,34], are the other approaches selected in this application task.

Recent studies have found that deep learning-based acoustic pattern recognition approaches [33,34], such as the DNN [35,36] and convolution neural network (CNN) [37,38] models, with MFCC features provide suitable performance in the dysarthric SCR task. More specifically, one-dimensional waveform signals were preprocessed by the MFCC feature extraction unit to obtain the two-dimensional spectrographic images used to train the CNN model. Following this, the trained CNN model was used to predict results from these two-dimensional spectrographic images in the application phase. Currently, this approach,

called CNN–MFCC in this study, is widely used in speech or acoustic event detection tasks. For example, Chen et al. [39] used the CNN–MFCC structure to predict the tones of Mandarin from input speech, and the results showed that this approach provided higher accuracy than classical approaches. Rubin et al. [40] applied the CNN–MFCC structure for the automatic classification of heart sounds, and the results suggested that this structure can also provide suitable performance in this application task. Che et al. [41] used a similar concept to CNN–MFCC in a partial discharge recognition task, and the results showed that the MFCC and CNN may be a promising event recognition method for this application too. The structure of CNN–MFCC be used to help dysarthric patients. Nakashika et al. [42] proposed a robust feature extraction method using a CNN model, which extracted the disordered speech features from a segment MFCC map. The experiment results of this study showed that the CNN-based feature extraction from the MFCC map provided better word-recognition results than other conventional feature extraction methods. More recently, Yakoub et al. [43] proposed an empirical model decomposition and Hurst-based model selection (EMDH)-CNN system to improve the recognition of dysarthric speech. The results showed that the proposed system provided higher accuracy than the hidden Markov with Gaussian Mixture model and the CNN model by 20.72% and 9.95%, respectively. From the above studies, we infer that a robust speech feature can benefit the acoustic pattern recognition system in the application of a dysarthric patient SCR task.

Recently, a novel speech feature, the phonetic posteriorgram (PPG), was proposed; this is a time-versus-class vector that expresses the posterior probabilities of phonetic classes for a specific timeframe (for a detailed description, refer to 2.2). Many studies' results show that the PPG feature can benefit many speech signal-processing tasks. For example, Zhao et al. [44] used PPG to process accent conversion, and the results showed a 20% improvement in speech quality. Zhou et al. [45] applied PPG to achieve cross-lingual voice conversion; the results showed effectiveness in intralingual and cross-lingual voice conversion between English and Mandarin speakers. More recently, in our previous study, the PPG was used to assist the gated CNN-based voice conversion model to convert dysarthric to normal speech [46], and the results showed that the PPG speech feature can benefit the voice conversion system for the dysarthric patient speech conversion task. Following the success of PPG features in previous dysarthric patient speech signal processing tasks, the first purpose of this study is to propose a hybrid system, called CNN–PPG, which is a CNN model with PPG features which could be used to improve the SCR accuracy for severe dysarthric patients. It should be noted that the goal of the proposed CNN–PPG in this study is to achieve high accuracy and stable recognition performance. Therefore, the concept of the personalized SCR system is also adopted in this study. The second purpose of this study is to compare the proposed CNN–PPG system with two classical systems (CNN model with MFCC features and ASR-based system) to ensure the benefits of this proposed system in this task. The third purpose of this study is to study the relation between the number of parameters and accuracy in these three systems, which can help us to reduce the cost of implementation in the future.

The rest of the article is organized as follows. Sections 2 and 3 presents the method and experimental results. Finally, Section 4 summarizes our findings.

2. Method

2.1. Material

We invited three CP patients to record 19 Mandarin commands 10 times each (seven times for the training set and the other three times for the testing set); the duration of each speech command was approximately one second, and the sampling rate was 16,000 Hz. These 19 commands included 10 action commands—close, up, down, previous, next, in, out, left, right, and home—and nine select commands—one, two, three, four, five, six, seven, eight, and nine—which were designed to allow dysarthric patients to control the proposed web browser app through their speech. These original data can be downloaded from the website <https://reurl.cc/a5vAG4> (accessed on 18 January 2021). To obtain more training

data for deep learning model training, the data augmentation approach [47] was used to obtain 103 (100 corruption with noise data and 3 time domain variance) augmentation data from the training set; meanwhile, the remaining three commands served as a testing set, without data augmentation. More specifically, we randomly selected 7 commands with the data augmentation approach to obtain the training set, and the other 3 commands were used as the testing set.

2.2. The Proposed CNN–PPG SCR System

Figure 1A shows the proposed CNN–PPG SCR system in this study, included training and testing phases. In the training phase, the speech commands of dysarthric patients (x_i) and the corresponding label result (t_i) were used to train the CNN model [48–50], where i was the frame index. The detailed structure of the CNN model used in the CNN–PPG SCR system is shown in Appendix A (Table A1), and achieved the best performance in this study. The x_i was processed by the unit of feature extraction (MFCC) to obtain the X_i^{MFCC} (i.e., 120-dimensional), first. Next, the unit of feature extraction (PPG) was used to convert the X_i^{MFCC} to a 33-dimensional PPG feature (X_i^{PPG}); for the detailed 33-dimensional phone, refer to Appendix B (Table A2). More specifically, the PPG features were obtained from the acoustic model of a speaker-dependent ASR system in Figure 1C, in which the TDNN [51,52] structure was used. A previous study, which used 40 coefficients, showed that the TDNN structure can effectively learn the temporal dynamics of speech signals [52]; hence, it could provide more benefits than alternative approaches to handle the issue of phonetic variation of a dysarthric patient's speech. The detailed training method of the acoustic model is described in the following section on the ASR-based SCR system. The parameters (φ_{CNN}) of the CNN model were trained and based on the PPG feature per frame (X_i^{PPG}) and the command probability vector (Y_i). More specifically, a nonlinear transfer function $F_{CNN}(\cdot)$ was learned to predict a probability vector (P_i) by X_i^{PPG} ; note that P_i should approach Y_i after the CNN model is well trained. The above description can be depicted as the following Equations (1) and (2).

$$F_{CNN}(X_i^{PPG} | \varphi_{CNN}) = f_h(\cdots f_3(f_2(f_1(X_i^{PPG} | \varphi_1) | \varphi_2) | \varphi_3) \cdots | \varphi_h) \quad (1)$$

$$F_{CNN}(X_i^{PPG} | \varphi_{CNN}) = P_i \simeq Y_i \quad (2)$$

Each operation $f_h(\cdot | \varphi_h)$ is defined as the h^{th} layer of the network. In addition, the convolution layer can be expressed as:

$$X_{h+1} = f_h(X_h | \varphi_h) = \text{relu}(W_h \otimes X_h + b_h), \quad (3)$$

$$\varphi_h = [W_h, b_h] \quad (4)$$

Note that W_h and b_h are defined as the collection of the h^{th} hidden layer's kernels (so-called filters) and the vector bias term, respectively. $\otimes$ represents a valid convolution, $\text{relu}(\cdot)$ is a piecewise linear activation function, and X_h is an output vector from the h^{th} layer. In this CNN-based SCR system, we applied a fully connective layer with softmax activation function as the final output [53]. The model uses cross entropy as the objective function ($L(\cdot)$) to adjust the parameter's (φ_{CNN}), objective function, as shown in Equation (5).

$$L(\varphi_{CNN}) = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^{19} y_{i,c} \log(p_{i,c}) \quad (5)$$

Figure 1. Three well-known speech command recognition (SCR) systems: (A) convolution neural network with Mel-frequency cepstral coefficient (CNN-MFCC), (B) convolution neural network with a phonetic posteriorgram (CNN-PPG) and (C) automatic speech recognition (ASR)-based.

In Equation (5), N represents the total number of frames used during training. $y_{i,c}$ is an element from the Y_i vector, which represents the ground truth probability value of the i^{th} frame being the c^{th} type, and $p_{i,c}$ comes from the P_i vector and represents the predicted scalar probability of the i^{th} frame being the c^{th} type. For model training, the back-propagation algorithm is used to find the suitable ϕ_{CNN} through the following equation:

$$\phi'_{CNN} = \arg\phi\min(L(\phi_{CNN})) \quad (6)$$

Further details of the CNN can be found in the descriptions in [48–50]. In the application phase, the trained CNN-PPG system was used to predict the results (t'_i) from the dysarthric patients' speech commands (x'_i), directly.

2.3. The Classical SCR Systems

2.3.1. CNN-MFCC SCR System

The block diagram of the CNN-MFCC model [39–41], which is the CNN model with MFCC features, is shown in Figure 1B. MFCC is a well-known feature extraction method [54] that has many successful applications in acoustic signal processing tasks [8,26,54–56]. This study applied the MFCC method to extract the acoustic features for the CNN model, as one SCR system used. The speech signals of dysarthric speech (x_i) were collected and labeled. The procedure for MFCC included six steps: pre-emphasis, windowing, fast Fourier transform, Mel-scale filter bank, nonlinear transformation, and discrete cosine transform. The high-frequency power of the speech signal declines during translation, which causes loss of high-frequency information. First, pre-emphasis is applied to compensate for the high-frequency signal [57]. Then, the frame-blocking unit is used to obtain the small frames from the input speech information ($s[n]$). Next, the Hamming windows method is used to alleviate the issue of side lobe plummets between each frame. Then, fast Fourier transform is applied to obtain the frequency response of each frame for spectral analysis; meanwhile, the Triangular Bandpass Filters (TBF) [58]

are used to integrate the frequency compositions from a Mel-filter band into one-energy intensity. Finally, the MFCC features ($c_l[n]$) are obtained based on the discrete cosine transform. Note that the librosa library [59] was used to obtain the X_i^{MFCC} in this study; 120-dimensional MFCC features (40-dimensional original MFCC + 40-dimensional velocity + 40-dimensional acceleration features) were used, because this setting provided suitable performance in previous studies [57,60,61].

Next, the obtained $c_l[n]$ and label target were used as the input and output of the CNN model, respectively. The CNN [48–50] model was applied in the CNN–MFCC SCR, for which the detailed structure is shown in Appendix C (Table A3). Next, the X_i^{MFCC} and relative label information were used as the input and output of the CNN model to learn the suitable parameters to identify the speech command of dysarthric patients. Note that the training approach for the CNN model is like the description in Equations (1)–(6) in Section 2.2. Finally, the trained CNN–MFCC model was used to predict the input command (x'_i) of dysarthric patients in the application phase.

2.3.2. ASR-Based SCR System

Figure 1C shows the ASR-based SCR system, which has three major parts: feature extraction, acoustic model, and language model [62]. First, the speech commands of dysarthric patients (x_i) were processed through the Kaldi feature extraction toolkit to obtain the 120-dimensional MFCC features (X_i^{MFCC}). Note that frame size and frame shift were set to 25 and 10 ms, respectively. After that, the X_i^{MFCC} was used as the input feature, and the related target, 33 phones, shown in Appendix B (Table A2), as an output to train this acoustic model, was used with the TDNN model. TDNN is a feed-forward architecture that can deal with speech signal context information through a designed hierarchical structure and layered process method from narrow context to long context information speech signals [52]. For learning wider temporal relationships, TDNN processes the input signals by slicing the output of the hidden activations from the previous layer through deeper layers to obtain important information [51]. The TDNN structure of the ASR system used in this study is shown in Appendix D (Table A4). In this study, we used the Kaldi toolkit [63] to train the ASR-based SCR system based on 41,097 (=3 patients $\times$ 19 commands $\times$ 7 times $\times$ 103 data augmentation) commands. The trained acoustic model will convert the X_i^{MFCC} to X_i^{PPG} ; meanwhile, the X_i^{PPG} will be used to train the language model in the ASR system. The language model is an important element of ASR, giving the probability of the next words through speech command data training. Traditionally, the language model of ASR uses an N-gram counting method to achieve its language model [64,65]. Recently, RNN has been applied to replace the N-gram structure [51] in complex tasks, such as multilanguage transfer. However, considering that RNN requires a long training time for similar accuracy performance, the HMM structure using the N-gram method was applied in this study as a language model. Next, the (X_i^{PPG}) features were used to train the language model. The detailed training approach for the ASR system refers to the Peddinti study [52]. Finally, the trained ASR-based SCR system could be used to predict the input commands (x'_i) of dysarthric patients, in the application phase.

2.4. Experiment Design

In this study, we proposed the CNN–PPG SCR system to help severely dysarthric patients control software via their speech commands, with two well-known SCR systems (CNN–MFCC and ASR-based SCR systems) used for comparison. First, the training set (described in Section 2.1) was used to train these three SCR systems. The speech commands of the training set were converted to MFCC features (X_i^{MFCC}) through the toolkit of Kaldi [63]. Next, the X_i^{MFCC} and related label targets were used as the input and output of the CNN–MFCC and ASR-based SCR systems to learn the parameters; the detailed training approach is described in Section 2.3.1 and Section 2.3.2, respectively. Meanwhile, the X_i^{MFCC} with labeled 33 class phone data for Mandarin (Wade–Giles system) were used to train the acoustic model, which is described in Section 2.3.2. Following this,

the trained acoustic model converted the MFCC features to PPG features (X_i^{PPG}), which were used to train the CNN model of the proposed CNN-PPG SCR system; the detailed setting of CNN-PPG is described in Section 2.2. Next, the testing set was used to test each SCR system to ensure the performance. In this study, the experiment was repeated 10 times, and the average results were used to compare the performance of each SCR system.

In order to ensure the benefits of the proposed CNN-PPG SCR system in this study, a two-experiment setup was used to investigate performance. First, we wanted to evaluate the benefits of the PPG features in the dysarthric patients' SCR tasks, with the MFCC features used as the comparison. The *t*-distributed stochastic neighbor embedding (*t*-SNE) [66] method was used to further compare them. Note that *t*-SNE technology is a machine learning algorithm for dimensionality reduction, based on a nonlinear dimensionality reduction technique. In this test, we used MFCC and PPG features to extract speech features for all data, and these extracted features were input to the *t*-SNE software for analysis. Next, we evaluated the performance of the proposed CNN-PPG SCR system with the CNN-MFCC and ASR-based SCR systems by the recognition rate method, with the ten times validation test approach used to confirm the performance of the three systems. We searched for the best setting by changing the model setting in this task (for details of the model setting, refer to Appendix E (Table A5), Appendix F (Table A6), and Appendix G (Table A7)); the best performance would represent the performance of each SCR system.

3. Results and Discussion

3.1. The Analysis of Speech Features between MFCC and PPG

Figure 2 shows the performance of these two speech features through the *t*-SNE approach, and the results indicated that these 19 commands of MFCC showed divergence and overlap in most frames; in contrast, the PPG feature showed more convergence and less overlapping frames in the same test condition. Specifically, from the *t*-SNE results, the PPG achieved more robust performance than MFCC to extract dysarthric speech features. Therefore, these results imply that the PPG could help the deep learning model achieve better performance in the dysarthric patient SCR task.

Figure 2. The *t*-SNE analysis of (A) CNN-MFCC model and (B) CNN-PPG model. In the CNN-PPG groups, the *t*-SNE of 19 commands were separate, distinct from the CNN-MFCC groups. Note: the labels “01” to “19” are the commands of close, up, down, previous, next, in, out, left, right, home, one, two, three, four, five, six seven, eight, and nine, respectively.

3.2. Recognition Performance of Each SCR System

Figure 3 shows the accuracy of the three models with different model sizes (related to the parameter numbers). The highest accuracy of the CNN-MFCC system was 65.67% (i.e., total parameter number 917,173), while the parameter numbers ranged from 471,169 to 917,173 and showed similar accuracy. The second model, the CNN-PPG system, showed

the best accuracy performance at 470,039 parameters, achieving 93.49%, while parameter numbers from 469,256 to 470,387 showed similar accuracy. The third model, ASR, showed the best performance at 89.59% with the parameter number 427,184, and the parameter numbers from 237,104 to 509,360 showed similar accuracy. Figure 4 shows the average recognition rate of a test repeated 10 times in different models: CNN-MFCC, CNN-PPG, and ASR-based. The average accuracy $\pm$ standard deviation of the CNN-MFCC system was $65.67\% \pm 3.9\%$, that of the CNN-PPG system was $93.49\% \pm 2.4\%$, and that of the ASR-based system was $89.63\% \pm 5.9\%$. The results indicate that CNN-PPG had the highest accuracy score and more stable performance compared with the other two systems. Hence, the PPG features are more robust than MFCC features in the CNN deep learning model. PPG features achieved higher recognition performance than MFCC in this study, which is consistent with previous studies [46]. The PPG features extract phone probability for each input frame; meanwhile, the monophone was used in the study. Therefore, it provided better performance than the well-known MFCC features. In addition, the results of this study show that the ASR-based SCR system provided better recognition performance than CNN-MFCC, though slightly lower than the CNN-PPG method. Thus, while previous studies indicated that the ASR-based SCR system is a gold standard in the application of the SCR task for dysarthric patients [21], the ASR system did not provide the best performance in this study. Furthermore, we investigated the cross-validation results ten times in each model. For the ASR-based SCR system, the best accuracy was 96.4%, in contrast to the best accuracy of CNN-PPG during ten times cross-validation history, 97.6%. The best accuracy of the two models was similar, but that of the proposed CNN-PPG was higher, due to the limited training data to decrease the ability of the ASR-based SCR system. Therefore, while on average, the ASR-based SCR system did not provide the best performance in this task, if a larger training set was available, the ASR-based system might provide better performance.

Figure 3. Comparison of difference model size of the three used models. The x-axis denotes the parameter numbers, and the y-axis denotes the recognition rate of speech command from dysarthric patients.

Figure 4. The average speech recognition rate (%) of CNN-MFCC, CNN-PPG, and ASR-based SCR systems in a test repeated 10 times.

Table 1 details the 10 experimental results (i.e., 10 repetitions/times) of these three SCR systems in the training and application phases. Note that each experiment was independent and separated and repeated 10 times. From the results of Table 1, we observed that there were no overfit [67,68] issues in the proposed CNN-PPG, because the performance between training and application phases was similar. In contrast, the CNN-MFCC system had the issue of overfit for all test times, with over 30% error between training and application phases. For the ASR-based system, there were also sometimes overfit issues in this study. These results indicated that the proposed CNN-PPG system can provide more stable performance than baseline systems for dysarthric patients in real application.

Table 1. Results of the 10 repeated experiments for each of the three systems.

Times	Accuracy (%)					
	Convolution Neural Network with Mel-frequency Cepstral Coefficient (CNN-MFCC)		Convolution Neural Network with a Phonetic Posteriorgram (CNN-PPG)		Automatic Speech Recognition (ASR)	
	Training Phase	Application Phase	Training Phase	Application Phase	Training Phase	Application Phase
1	97.9%	57.9%	95.4%	95.3%	100%	89.5%
2	98.2%	67.3%	95.7%	94.2%	100%	94.2%
3	98.2%	63.7%	93.2%	95.3%	100%	89.5%
4	97.9%	67.8%	96.7%	96.5%	100%	74.9%
5	96.7%	69.6%	95.7%	93.6%	100%	87.1%
6	92.2%	71.9%	96.9%	92.9%	100%	94.2%
7	95.9%	64.9%	95.2%	90.0%	99.7%	94.2%
8	99.2%	64.9%	98.9%	90.0%	100%	88.9%
9	97.9%	62.0%	97.2%	91.2%	100%	95.2%
10	96.7%	66.7%	96.2%	95.9%	100%	88.9%
Average	97.1%	65.7%	96.1%	93.4%	99.9%	89.6%

The number of parameters of the system used is related to the implementation costs, such as computation consumption and memory size; meanwhile, a large model size (i.e., higher parameter numbers) causes longer response time and has battery lifetime issues [69,70]. Therefore, under a similar recognition performance condition, a smaller model size could provide more benefits for users. The experimental results show that the CNN-PPG can provide a higher recognition rate than the other two SCR systems with a lower model size than the ASR-based system. Hence, these results suggest that the CNN-PPG system can provide practical feasibility in the future.

3.3. The Existing Application of Deep Learning Technology in Healthcare

Recently, many medical and healthcare devices based on deep learning technology have been proposed for tasks such as the detection of pathological voice [71], healthcare monitoring [72], heart disease prediction [73], detection and reconstruction of dysarthric speech [74], and speech waveform synthesis [75]. Through the application of deep learning with big data, we gain many benefits for healthcare applications compared with traditional approaches. For the deep learning approach, the training data are among the most important parts. Therefore, how to efficiently obtain useful training materials will be a very important matter in the future employment of deep learning technology in medical care applications.

4. Conclusions and Future Works

This study aimed to use a deep learning-based SCR system to help dysarthric patients control mobile devices via speech. Our results showed that: (1) the speech feature of PPG can achieve better recognition performance than MFCC and (2) the proposed CNN-PPG SCR system can provide higher recognition accuracy than two classical SCR systems in this study; meanwhile, only a small model size is needed to compare the CNN-MFCC and ASR-based SCR systems. More specifically, the average accuracy of CNN-PPG achieved an acceptable performance (i.e., 93.49% recognition rate) in this study; therefore, the CNN-PPG can be applied in the SCR system to help dysarthric patients control a communication device using their speech commands, as shown in Figure 5. Specifically, the dysarthric patients can use combinations of 19 commands to select application software functions (e.g., YouTube, Facebook, and messaging). In addition, we plan to use natural language processing technology to provide automated options from the interlocutor's speech; meanwhile, these candidates' response sentences will be selected through the patient's 19 commands, thereby accelerating the patient's response rate. Finally, we plan to implement this proposed CNN-PPG in a mobile device to help dysarthric patients improve their communication quality in future studies.

Figure 5. An example of the processed SCR system for dysarthric speakers. Users can maneuver through number commands to select the program they want to use or the help they need. The green area is the item that can be controlled by voice command.

The proposed CNN–PPG system provided higher performance than two baseline SCR systems in a personalized application condition; however, the performance of the proposed system could be limited in general application (i.e., without asking the user to record the speech commands), because of the challenging the issues of individual variability and phonetic variation in the speech [10–12] of dysarthric patients. To overcome this challenge and further improve system benefits, two possible research directions appear. First, we will provide this system for free use to dysarthric patients and many patients’ voice commands, with the users’ consent. Next, the data obtained can, therefore, be used to retrain the CNN–PPG system to improve performance in a generalization application condition. Second, advanced deep learning technologies [50,76–79], such as few-shots audio classification, 1D/2D CNN audio temporal integration, semi-supervised speaker diarization deep neural embeddings, convolutional 2D audio stream management, and data augmentation, can be used to further improve the performance of the proposed CNN–PPG SCR system in future studies.

Author Contributions: Conceptualization, Y.-H.L., G.-M.H., C.-Y.C., and Y.-Y.L.; Methodology, Y.-H.L.; Software, W.-Z.Z., W.-Z.Z., and Y.-H.H.; Validation, Y.-Y.L., J.-Y.H., and Y.-H.L.; Formal Analysis, W.-Z.Z.; Investigation, Y.-Y.L. and Y.-H.L.; Resources, G.-M.H., C.-Y.C., and Y.-H.L.; Data Curation, W.-Z.Z.; Writing—Original Draft Preparation, Y.-H.L. and Y.-Y.L.; Writing—Review and Editing, Y.-Y.L. and Y.-H.L.; Visualization, Y.-H.H.; Supervision, Y.-H.L.; Project Administration, J.-Y.H. and Y.-H.L.; Funding Acquisition, Y.-H.L., G.-M.H., and C.-Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by funds awarded by the Ministry of Science and Technology, Taiwan, under Grant MOST 109-2218-E-010-004, and by industry–academia cooperation project funding from APrevent Medical (107J042).

Institutional Review Board Statement: This investigation was approved by the Ethics Committee of Taipei Medical University—Joint Institutional Board (N201607030) and was performed according to the principles and policies of the Declaration of Helsinki.

Informed Consent Statement: Informed written consent for participation was gained from all participants.

Data Availability Statement: Data are presented in the manuscript; further information is available upon request data.

Acknowledgments: This study was supported by the Ministry of Science of Technology of Taiwan (109-2218-E-010-004) and an industry–academia cooperation project of APrevent Medical Inc. (107J042).

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

Appendix A. The Setting of CNN–PPG System

Table A1. Structure of the CNN–PPG-based SCR system.

Input: 120 D, Output: 19 Class	
Hidden Layer	
Layer 1	filters: 10, kernel size: 3×3 , strides: 2×2 , ReLU
Layer 2	filters: 8, kernel size: 3×3 , strides: 2×2 , ReLU
Layer 3	filters: 10, kernel size: 3×3 , strides: 2×2 , ReLU
Global average pooling, Dense (19), softmax	

Appendix B. The 33-Dimensional Data of PPG

Table A2. The 33 phones applied in this study and correlated class index.

Class Index	Phone	Class Index	Phone	Class Index	Phone
1	SIL	12	h	23	o1
2	a1	13	i1	24	o3
3	a2	14	i2	25	o4
4	a3	15	i3	26	q
5	a4	16	i4	27	s
6	b6	17	ii4	28	sh
7	d7	18	j	29	u1
8	e4	19	l	30	u3
9	err4	20	ng4	31	u4
10	f	21	nn1	32	x
11	g	22	nn2	33	z

Appendix C. The Setting of CNN–MFCC System

Table A3. Structure of the CNN–MFCC system.

Input: 120 D, Output: 19 Class	
Layer 1	filters: 12, kernel size: 3×3 , strides: 2×2 , PReLU
Layer 2	filters: 12, kernel size: 3×3 , strides: 2×2 , PReLU
Layer 3	filters: 24, kernel size: 3×3 , strides: 2×2 , PReLU
Layer 4	filters: 24, kernel size: 3×3 , strides: 1×1 , PReLU
Layer 5	filters: 48, kernel size: 3×3 , strides: 1×1 , PReLU
Layer 6	filters: 48, kernel size: 3×3 , strides: 1×1 , PReLU
Layer 7	filters: 96, kernel size: 3×3 , strides: 1×1 , PReLU
Layer 8	filters: 96, kernel size: 3×3 , strides: 1×1 , PReLU
Layer 9	filters: 192, kernel size: 3×3 , strides: 1×1 , PReLU, dropout (0.4)
Layer 10	filters: 192, kernel size: 3×3 , strides: 1×1 , PReLU, dropout (0.3)
Global average pooling, Dense (19), Dropout (0.2), softmax	

Note: the three convolution layers structure could be used in this model, which achieved the best performance in this task; meanwhile, we apply global average pooling [80] and a fully connected layer after three convolution layers in this model.

Appendix D. The Setting of ASR-Based Model

Table A4. Structure of the Time Delay Neural Network (TDNN) applied in ASR-based model.

Input: 120 D; Output: 33 Class	
Hidden Layer	
Layer 1	dims: 128, context_size = 3, dilation = 14, ReLU
Layer 2	dims: 128, context_size = 3, dilation = 14, ReLU
Layer 3	dims: 128, context_size = 3, dilation = 14, ReLU
Layer 4	dims: 128, context_size = 3, dilation = 14, ReLU
Layer 5	dims: 128, context_size = 3, dilation = 14, ReLU
Layer 6	dims: 128, context_size = 3, dilation = 14, ReLU
Layer 7	dims: 128, context_size = 3, dilation = 14, ReLU
Layer 8	dims: 128, context_size = 3, dilation = 14, ReLU
Layer 9	dims: 128, context_size = 3, dilation = 14, ReLU
Layer 10	dims: 128, context_size = 3, dilation = 14, ReLU
Layer 11	dims: 128, context_size = 3, dilation = 14, ReLU
Dense (33), softmax	

Appendix E. The Layers and Parameter Number Setting of CNN–MFCC System

Table A5. Filter numbers of different layers in CNN–MFCC system.

	Model Size					
	A	B	C	D	E	F
layer1	8	9	10	12	14	16
layer2	8	9	10	12	14	16
layer3	16	18	20	24	28	32
layer4	16	18	20	24	28	32
layer5	32	36	40	48	56	64
layer6	32	36	40	48	56	64
layer7	64	72	80	96	112	128
layer8	64	72	80	96	112	128
layer9	128	144	160	192	224	256
layer10	128	144	160	192	224	256
output	19	19	19	19	19	19
Total	303,355	382,663	471,169	675,775	917,173	1,195,363

Appendix F. The Layers and Parameter Number Setting of CNN–PPG System

Table A6. Filter numbers of different layers in CNN–PPG SCR system.

	Model Size						
	A	B	C	D	E	F	G
layer1	5	6	7	8	9	10	12
layer2	3	4	5	5	6	8	8
layer3	5	6	7	8	9	10	12
output	19	19	19	19	19	19	19
Total	442	635	864	984	1267	1767	2115

Appendix G. The Layers and Parameter Number Setting of ASR-Based SCR System

Table A7. The different settings of ASR-based SCR system.

	Model Size							
	A	B	C	D	E	F	F	H
layer1	128	128	128	128	128	128	128	128
layer2	128	128	128	128	128	128	128	128
layer3	128	128	128	128	128	128	128	128
layer4	128	128	128	128	128	128	128	128
layer5	0	128	128	128	128	128	128	128
layer6	0	0	128	128	128	128	128	128
layer7	0	0	0	128	128	128	128	128
layer8	0	0	0	0	128	128	128	128
layer9	0	0	0	0	0	128	128	128
layer10	0	0	0	0	0	0	128	128
layer11	0	0	0	0	0	0	0	128
output	33	33	33	33	33	33	33	33
Total	237,104	278,192	319,280	360,368	396,336	427,184	468,272	509,360

References

1. Darley, F.L.; Aronson, A.E.; Brown, J.R. Differential Diagnostic Patterns of Dysarthria. *J. Speech Hear. Res.* **1969**, *12*, 246–269. [[CrossRef](#)] [[PubMed](#)]
2. Calculator, S.; Luchko, C.D. Evaluating the Effectiveness of a Communication Board Training Program. *J. Speech Hear. Disord.* **1983**, *48*, 185–191. [[CrossRef](#)]
3. Birchfield, S. Elliptical head tracking using intensity gradients and color histograms. In Proceedings of the Proceedings 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231), Santa Barbara, CA, USA, 25 June 1998; IEEE: New York, NY, USA, 2002.
4. Zhou, Q.; Xing, J.; Chen, W.; Zhang, X.; Yang, Q. From Signal to Image: Enabling Fine-Grained Gesture Recognition with Commercial Wi-Fi Devices. *Sensors* **2018**, *18*, 3142. [[CrossRef](#)] [[PubMed](#)]
5. Lin, C.-S.; Ho, C.-W.; Chen, W.-C.; Chiu, C.-C.; Yeh, M.-S. Powered wheelchair controlled by eye-tracking system. *Opt. Appl.* **2006**, *36*, 401–412.
6. Trnka, K.; McCaw, J.; Yarrington, D.; McCoy, K.F.; Pennington, C. Word prediction and communication rate in AAC. In Proceedings of the IASTED International Conference on Telehealth/ Assistive Technologies, Baltimore, MD, USA, 16–18 April 2008.
7. Rosen, K.; Yampolsky, S. Automatic speech recognition and a review of its functioning with dysarthric speech. *Augment. Altern. Commun.* **2000**, *16*, 48–60. [[CrossRef](#)]
8. Shahamiri, S.R.; Salim, S.S.B. Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach. *Adv. Eng. Inform.* **2014**, *28*, 102–110. [[CrossRef](#)]
9. Sharma, H.V.; Hasegawa-Johnson, M. Acoustic model adaptation using in-domain background models for dysarthric speech recognition. *Comput. Speech Lang.* **2013**, *27*, 1147–1162. [[CrossRef](#)]
10. Carrillo, L.; Ortiz, K.Z. Vocal analysis (auditory-perceptual and acoustic) in dysarthrias. *Pró-Fono Rev. Atualização Científica* **2007**, *19*, 381–386. [[CrossRef](#)]
11. Le Dorze, G.; Ouellet, L.; Ryalls, J. Intonation and speech rate in dysarthric speech. *J. Commun. Disord.* **1994**, *27*, 1–18. [[CrossRef](#)]
12. Weismer, G.; Tjaden, K.; Kent, R.D. Can articulatory behavior in motor speech disorders be accounted for by theories of normal speech production? *J. Phon.* **1995**, *23*, 149–164. [[CrossRef](#)]
13. O'Sullivan, S.B.; Schmitz, T.J. *Physical Rehabilitation*, 5th ed.; Davis Company: Philadelphia, PA, USA, 2007.
14. Hasegawa-Johnson, M.; Gunderson, J.; Perlman, A.; Huang, T. HMM-Based and SVM-Based Recognition of the Speech of Talkers with Spastic Dysarthria. In Proceedings of the 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, France, 14–19 May 2006; IEEE: New York, NY, USA, 2006.
15. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [[CrossRef](#)]
16. Rudzicz, F. Phonological features in discriminative classification of dysarthric speech. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; IEEE: New York, NY, USA, 2009; pp. 4605–4608.
17. Rudzicz, F. Articulatory Knowledge in the Recognition of Dysarthric Speech. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *19*, 947–960. [[CrossRef](#)]
18. Vázquez, J.J.; Arjona, J.; Linares, M.; Casanovas-Garcia, J. A Comparison of Deep Learning Methods for Urban Traffic Forecasting using Floating Car Data. *Transp. Res. Procedia* **2020**, *47*, 195–202. [[CrossRef](#)]
19. Nguyen, G.; Dlugolinsky, S.; Tran, V.; Garcia, A.L. Deep Learning for Proactive Network Monitoring and Security Protection. *IEEE Access* **2020**, *8*, 19696–19716. [[CrossRef](#)]
20. Carta, S.; Ferreira, A.; Podda, A.S.; Recupero, D.R.; Sanna, A. Multi-DQN: An ensemble of Deep Q-learning agents for stock market forecasting. *Expert Syst. Appl.* **2021**, *164*, 113820. [[CrossRef](#)]
21. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.-C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019.
22. Fathima, N.; Patel, T.; Mahima, C.; Iyengar, A. TDNN-based Multilingual Speech Recognition System for Low Resource Indian Languages. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018.
23. Hawley, M.S.; Enderby, P.; Green, P.; Cunningham, S.; Brownsell, S.; Carmichael, J.; Parker, M.; Hatzis, A.; O'Neill, P.; Palmer, R. A speech-controlled environmental control system for people with severe dysarthria. *Med Eng. Phys.* **2007**, *29*, 586–593. [[CrossRef](#)]
24. Fager, S.K.; Beukelman, D.R.; Jakobs, T.; Hosom, J.-P. Evaluation of a Speech Recognition Prototype for Speakers with Moderate and Severe Dysarthria: A Preliminary Report. *Augment. Altern. Commun.* **2010**, *26*, 267–277. [[CrossRef](#)]
25. Yang, S.; Chung, M. Improving Dysarthric Speech Intelligibility using Cycle-consistent Adversarial Training. In Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies 2020, Valetta, Malta, 24–26 February 2020.
26. Selouani, S.-A.; Yakoub, M.S.; O'Shaughnessy, D. Alternative Speech Communication System for Persons with Severe Speech Disorders. *EURASIP J. Adv. Signal Process.* **2009**, *2009*, 540409. [[CrossRef](#)]
27. Polur, P.D.; Miller, G.E. Investigation of an HMM/ANN hybrid structure in pattern recognition application using cepstral analysis of dysarthric (distorted) speech signals. *Med Eng. Phys.* **2006**, *28*, 741–748. [[CrossRef](#)] [[PubMed](#)]

28. Hawley, M.; Enderby, P.; Green, P.; Brownsell, S.; Hatzis, A.; Parker, M.; Carmichael, J.; Cunningham, S.; O'Neill, P.; Palmer, R. STARDUST—Speech Training and Recognition for Dysarthric Users of Assistive Technology. In Proceedings of the 7th European Conference for the Advancement of Assistive Technology (AAATE 2003), Dublin, Ireland, 31 August–3 September 2003.
29. Farooq, O.; Datta, S. Mel filter-like admissible wavelet packet structure for speech recognition. *IEEE Signal Process. Lett.* **2001**, *8*, 196–198. [\[CrossRef\]](#)
30. Saia, R.; Carta, S.; Fenu, G. A Wavelet-based Data Analysis to Credit Scoring. In Proceedings of the 2nd International Conference on Cryptography, Security and Privacy, Association for Computing Machinery (ACM), Tokyo, Japan, 25–27 February 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 176–180.
31. Du, J.; Huo, Q. A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions. In Proceedings of the Ninth Annual Conference of the International Speech Communication Association, Brisbane, Australia, 22–26 September 2008.
32. Davis, S.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *28*, 357–366. [\[CrossRef\]](#)
33. Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J. API design for machine learning software: Experiences from the scikit-learn project. In Proceedings of the European Conference on Machine Learning and Principles of Knowledge Discovery in Databases, Prague, Czech Republic, 23–27 September 2013.
34. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
35. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.-R.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.N.; et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [\[CrossRef\]](#)
36. Yilmaz, E.; Ganzeboom, M.; Cucchiari, C.; Strik, H. Multi-Stage DNN Training for Automatic Recognition of Dysarthric Speech. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017.
37. Abdel-Hamid, O.; Mohamed, A.; Jiang, H.; Deng, L.; Penn, G.; Yu, D. Convolutional neural networks for speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1533–1545. [\[CrossRef\]](#)
38. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [\[CrossRef\]](#)
39. Chen, C.; Bunesco, R.; Xu, L.; Liu, C. Tone Classification in Mandarin Chinese Using Convolutional Neural Networks. In Proceedings of the Interspeech, San Francisco, CA, USA, 8–12 September 2016.
40. Rubin, J.; Abreu, R.; Ganguli, A.; Nelaturi, S.; Matei, I.; Sricharan, K. Classifying Heart Sound Recordings using Deep Convolutional Neural Networks and Mel:Frequency Cepstral Coefficients. In Proceedings of the 2016 Computing in Cardiology Conference (CinC), Computing in Cardiology, Vancouver, ON, Canada, 11–14 September 2016.
41. Che, Q.; Wen, H.; Li, X.; Peng, Z.; Chen, K.P. Partial Discharge Recognition Based on Optical Fiber Distributed Acoustic Sensing and a Convolutional Neural Network. *IEEE Access* **2019**, *7*, 101758–101764. [\[CrossRef\]](#)
42. Nakashika, T.; Yoshioka, T.; Takiguchi, T.; Ariki, Y.; Duffner, S.; Garcia, C. Convolutional Bottleneck Network with Dropout for Dysarthric Speech Recognition. *Trans. Mach. Learn. Artif. Intell.* **2014**, *2*, 48–62. [\[CrossRef\]](#)
43. Yakoub, M.S.; Selouani, S.-A.; Zaidi, B.-F.; Bouchair, A. Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network. *EURASIP J. Audio Speech Music. Process.* **2020**, *2020*, 1–7. [\[CrossRef\]](#)
44. Zhao, G.; Sonsaat, S.; Levis, J.; Chukharev-Hudilainen, E.; Gutierrez-Osuna, R. Accent Conversion Using Phonetic Posteriorgrams. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: New York, NY, USA, 2018; pp. 5314–5318.
45. Zhou, Y.; Tian, X.; Xu, H.; Das, R.K.; Li, H. Cross-lingual Voice Conversion with Bilingual Phonetic Posteriorgram and Average Modeling. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: New York, NY, USA, 2019; pp. 6790–6794.
46. Chen, C.Y.; Zheng, W.Z.; Wang, S.S.; Tsao, Y.; Li, P.C.; Lai, Y.H. Enhancing Intelligibility of Dysarthric Speech Using Gated Convolutional-based Voice Conversion System to appear. In Proceedings of the IEEE Interspeech, Shanghai, China, 25–29 October 2020.
47. Chen, J.; Wang, Y.; Yoho, S.E.; Wang, D.; Healy, E.W. Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises. *J. Acoust. Soc. Am.* **2016**, *139*, 2604–2612. [\[CrossRef\]](#)
48. Bengio, Y. Learning Deep Architectures for AI. *Found. Trends Mach. Learn.* **2009**, *2*, 1–127. [\[CrossRef\]](#)
49. Mohamed, A.; Dahl, G.E.; Hinton, G. Acoustic modeling using deep belief networks. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *20*, 14–22. [\[CrossRef\]](#)
50. Salamon, J.; Bello, J.P. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Process. Lett.* **2017**, *24*, 279–283. [\[CrossRef\]](#)
51. Garcia-Romero, D.; McCree, A. Stacked Long-Term TDNN for Spoken Language Recognition. In Proceedings of the Interspeech 2016, San Francisco, CA, USA, 8–12 September 2016. [\[CrossRef\]](#)
52. Peddinti, V.; Povey, D.; Khudanpur, S. A time delay neural network architecture for efficient modeling of long temporal contexts. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.

53. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *Mach. Learn.* **2015**, arXiv:1503.02531.
54. Teller, V. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. *Comput. Linguist.* **2000**, *26*, 638–641. [\[CrossRef\]](#)
55. Räsänen, O.; Leppänen, J.; Laine, U.K.; Saarinen, J.P. Comparison of classifiers in audio and acceleration based context classification in mobile phones. In Proceedings of the European Signal Processing Conference, Barcelona, Spain, 29 August–2 September 2011; IEEE: New York, NY, USA, 2011.
56. Lai, Y.-H.; Tsao, Y.; Lu, X.; Chen, F.; Su, Y.-T.; Chen, K.-C.; Chen, Y.-H.; Chen, L.-C.; Li, L.P.-H.; Lee, C.-H. Deep Learning-Based Noise Reduction Approach to Improve Speech Intelligibility for Cochlear Implant Recipients. *Ear Hear.* **2018**, *39*, 795–809. [\[CrossRef\]](#)
57. Tiwari, V. MFCC and its applications in speaker recognition. *Int. J. Emerg. Technol.* **2010**, *1*, 19–22.
58. Ganchev, T.; Fakotakis, N.; Kokkinakis, G. Comparative evaluation of various MFCC implementations on the speaker verification task. In Proceedings of the SPECOM, Patras, Greece, 17–19 October 2005.
59. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. librosa: Audio and music signal analysis in python. In Proceedings of the 14th Python in Science Conference, Austin, TX, USA, 6–12 July 2015.
60. Furui, S. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoust. Speech, Signal Process.* **1981**, *29*, 254–272. [\[CrossRef\]](#)
61. Ma, L.; Milner, B.; Smith, D. Acoustic environment classification. *ACM Trans. Speech Lang. Process.* **2006**, *3*, 1–22. [\[CrossRef\]](#)
62. Masmoudi, A.; Bougares, F.; Ellouze, M.; Estève, Y.; Belguith, L.H. Automatic speech recognition system for Tunisian dialect. *Lang. Resour. Evaluation* **2018**, *52*, 249–267. [\[CrossRef\]](#)
63. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi speech recognition toolkit. In Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Waikoloa, HI, USA, 11–15 December 2011.
64. Deoras, A.; Mikolov, T.; Kombrink, S.; Karafiat, M.; Khudanpur, S. Variational approximation of long-span language models for lvcsr. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; IEEE: New York, NY, USA, 2011; pp. 5532–5535.
65. Chen, S.F.; Goodman, J. An empirical study of smoothing techniques for language modeling. *Comput. Speech Lang.* **1999**, *13*, 359–394. [\[CrossRef\]](#)
66. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
67. Sarle, W.S. Stopped training and other remedies for overfitting. *Comput. Sci. Stat.* **1996**, *352*, 60.
68. Lawrence, S.; Giles, C.L.; Tsoi, A.C. Lessons in neural network training: Overfitting may be harder than expected. In Proceedings of the Fourteenth National Conference on Artificial Intelligence, Menlo Park, CA, USA, 14–19 July 1997.
69. Liu, Z.; Wang, Y.; Zhang, J.; Liu, Z. Shortcut computation for the thermal management of a large air-cooled battery pack. *Appl. Therm. Eng.* **2014**, *66*, 445–452. [\[CrossRef\]](#)
70. Xian, C.; Lu, Y.-H.; Li, Z. Adaptive computation offloading for energy conservation on battery-powered systems. In Proceedings of the 2007 International Conference on Parallel and Distributed Systems, Hsinchu, Taiwan, 5–7 December 2007; IEEE: New York, NY, USA, 2007; pp. 1–8.
71. Fang, S.-H.; Tsao, Y.; Hsiao, M.-J.; Chen, J.-Y.; Lai, Y.-H.; Lin, F.-C.; Wang, C.-T. Detection of Pathological Voice Using Cepstrum Vectors: A Deep Learning Approach. *J. Voice* **2019**, *33*, 634–641. [\[CrossRef\]](#)
72. Ali, F.; El-Sappagh, S.; Islam, S.M.R.; Ali, A.; Attique, M.; Imran, M.; Kwak, K.-S. An intelligent healthcare monitoring framework using wearable sensors and social networking data. *Future Gener. Comput. Syst.* **2020**, *114*, 23–43. [\[CrossRef\]](#)
73. Ali, F.; El-Sappagh, S.; Islam, S.R.; Kwak, D.; Ali, A.; Imran, M.; Kwak, K.-S. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Inf. Fusion* **2020**, *63*, 208–222. [\[CrossRef\]](#)
74. Korzekwa, D.; Barra-Chicote, R.; Kostek, B.; Drugman, T.; Lajszczak, M. Interpretable Deep Learning Model for the Detection and Reconstruction of Dysarthric Speech. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019.
75. Merritt, T.; Putrycz, B.; Nadolski, A.; Ye, T.; Korzekwa, D.; Dolecki, W.; Drugman, T.; Klimkov, V.; Moinet, A.; Breen, A.; et al. Comprehensive Evaluation of Statistical Speech Waveform Synthesis. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; IEEE: New York, NY, USA, 2018; pp. 325–331.
76. Vryzas, N.; Tsipas, N.; Dimoulas, C. Web Radio Automation for Audio Stream Management in the Era of Big Data. *Information* **2020**, *11*, 205. [\[CrossRef\]](#)
77. Vrysis, L.; Tsipas, N.; Thoidis, I.; Dimoulas, C. 1D/2D Deep CNNs vâs. Temporal Feature Integration for General Audio Classification. *J. Audio Eng. Soc.* **2020**, *68*, 66–77. [\[CrossRef\]](#)
78. Tsipas, N.; Vrysis, L.; Konstantoudakis, K.; Dimoulas, C. Semi-supervised audio-driven TV-news speaker diarization using deep neural embeddings. *J. Acoust. Soc. Am.* **2020**, *148*, 3751–3761. [\[CrossRef\]](#)
79. Brezeale, D.; Cook, D.J. Using closed captions and visual features to classify movies by genre. In Proceedings of the Poster Session of the Seventh International Workshop on Multimedia Data Mining, Philadelphia, PA, USA, 20 August 2006.
80. Lin, M.; Chen, Q.; Yan, S. Network in network. *Neural Evol. Comput.* **2013**, arXiv:1312.4400.