Interspeech 2018
2-6 September 2018, Hyderabad

Dysarthric Speech Recognition Using Convolutional LSTM Neural Network

Myungjong Kim[1], *Beiming Cao*[1], *Kwanghoon An*[1], *Jun Wang*[1,2]

[1]Speech Disorders & Technology Lab, Department of Bioengineering
[2]Callier Center for Communication Disorders, University of Texas at Dallas, United States

{myungjong.kim, beiming.cao, kwanghoon.an, wangjun}@utdallas.edu

Abstract

Dysarthria is a motor speech disorder that impedes the physical production of speech. Speech in patients with dysarthria is generally characterized by poor articulation, breathy voice, and monotonic intonation. Therefore, modeling the spectral and temporal characteristics of dysarthric speech is critical for better performance in dysarthric speech recognition. Convolutional long short-term memory recurrent neural networks (CLSTM-RNNs) have recently successfully been used in normal speech recognition, but have rarely been used in dysarthric speech recognition. We hypothesized CLSTM-RNNs have the potential to capture the distinct characteristics of dysarthric speech, taking advantage of convolutional neural networks (CNNs) for extracting effective local features and LSTM-RNNs for modeling temporal dependencies of the features. In this paper, we investigate the use of CLSTM-RNNs for dysarthric speech recognition. Experimental evaluation on a database collected from nine dysarthric patients showed that our approach provides substantial improvement over both standard CNN and LSTM-RNN based speech recognizers.

Index Terms: Dysarthria, convolutional neural network, long short-term memory recurrent neural network (LSTM-RNN), speech recognition

1. Introduction

Individuals with dysarthria, a neurological motor speech disorder, have trouble controlling their motor subsystems including respiration, phonation, resonance, articulation, and prosody [1]. Speech in patients with dysarthria is generally characterized by poor articulation, breathy voice, and monotonic intonation [1]. Therefore, standard automatic speech recognition (ASR) methods for the general public typically do not perform well for patients with dysarthria.

Related work on the recognition of dysarthric speech has been mostly focused on acoustic modeling to capture the acoustic cues of dysarthric speech. A variety of acoustic models such as Gaussian mixture model (GMM)-hidden Markov models (HMMs), support vector machine (SVM), and artificial neural networks were studied [2–4]. Also, deep neural network (DNN)-HMM based acoustic models were widely applied to dysarthric speech recognition [5, 6].

Convolutional neural networks (CNNs) have been sucessfully applied to automatic speech recognition due to the ability of extracting local features through convolution and pooling operations [7]. There are several types of CNNs, including frequency-domain CNNs (F-CNNs), time-domain CNNs (T-CNNs), and time-frequency CNNs (TF-CNNs). CNNs have been demonstrated effective in extracting useful features in spectral, temporal, and spectro-temporal domains that are robust to small variations by using convolution and pooling along the frequency axis, the time axis, and the time-frequency region, respectively. These models were successfully applied to speech recognition applications [8–10]. In particular, TF-CNN based bottleneck features were used for dysarthric speech recognition and the features were better than standard mel-frequency cepstral coefficients on GMM-HMM based ASR systems [11]. Recently, parallel TF-CNNs (PTF-CNNs), where separate F-CNN and T-CNN are combined with fully connected layers, were introduced for noise robust speech recognition [12].

Long short-term memory recurrent neural networks (LSTM-RNNs) can capture long-range temporal dependencies by overcoming the vanishing gradient problem in conventional recurrent neural networks (RNNs) [13]. It has been successfully used in speech recognition applications [14]. LSTM-RNNs were also applied to dysarthric speech recognition [15]. In [15], LSTM-RNNs produced better performance than standard DNN for most mildly dysarthric speakers while LSTM-RNNs gave worse performance for severely dysarthric speakers.

Combining CNNs and LSTM-RNNs, called convolutional LSTM-RNNs (CLSTM-RNNs), has benefit from CNNs for local feature extraction and LSTM-RNNs for temporal modeling, and it has shown better performance than CNNs alone and LSTM-RNNs alone for sequential modeling such as speech recognition [16, 17] and music tagging [18]. People with dysarthria have the problem in controlling speech muscles, and therefore, acoustic cues in time-frequency region are often shifted [11]. These speech characteristics are significantly dependent in time. Therefore, CLSTM-RNNs might be more effective in modeling dysarthric speech than CNNs or LSTM-RNNs alone. However, CLSTM-RNNs have rarely been studied in dysarthric speech recognition.

In this paper, we investigate the effectiveness of CLSTM-RNNs for the phoneme recognition of dysarthric speech. We tested four types of CNNs including F-CNNs, T-CNNs, TF-CNNs, and PTF-CNNs, and their combinations with LSTM-RNNs. Our approach was evaluated on a dataset of phrases collected from nine dysarthric speakers with multiple recording sessions in a speaker-independent way. The experimental results showed CLSTM-RNNs produce promising performance over either CNNs or LSTM-RNNs alone. In particular, the overall ASR performance was the best on time-frequency convolutional LSTM-RNNs. Further, we compared the performance across speakers and sessions.

2. Dysarthric Speech Data

A dysarthric speech dataset collected from nine patients with amyotrophic lateral sclerosis (ALS) (6 females and 3 males) was used. The participants were all American English talkers. ALS is also known as Lou Gehrig's disease, which is one of the most common motor neuron diseases [19], resulting in progressive degeneration of both upper and lower motor neurons [19]. Four of the patients visited the lab more than once for the data

10.21437/Interspeech.2018-2250

collection. The average duration between consecutive sessions is six months for those patients. Thus, we collected the speech data in eighteen sessions from nine patients in total.

The average age at their first visit was 62.8 years old (SD=8.8). They are all early diagnosed (within half to one year) but as ALS progresses the speech intelligibility gets worse. Thus, we measured perceptual speech intelligibility scores for each session. The speech intelligibility of those participants with ALS varied from normal (100%) to severely unintelligible speech (0%). Speech intelligibility is diagnosed by a speech language pathologist. To understand the performance of our speech recognition algorithms for different levels of dysarthria, we divided the data set into three groups based on their speech intelligibility: eleven sessions as high (above 90%), four sessions as middle (65-90%), and three sessions as low (below 65%).

During each recording session, each subject produced up to 4 repetitions of 20 unique sentences at their habitual speaking rate and loudness. These sentences are selected in daily conversations (e.g., *How are you doing?*) or related to patient's daily use (e.g., *I need to make an appointment.*). In total, 1,289 utterances for 20 unique phrases were collected. The number of phonemes is 17,712 and the number of unique phonemes is 39.

We also collected normal speech data from seven American English speakers (four females and three males). The mean age of the participants was 25.4 years old (SD=3.6). No history of speech, language, or cognitive problems from any participant was reported. Each subject participated in one session in which he or she repeated a list of 132 phrases twice at their habitual speaking rate. The phrases that are frequently used in daily life were selected from [20]. Fourteen phrases from the ALS dataset and normal dataset were overlapped. These normal speech data were used for acoustic model training, which will be discussed in Section 4.1. The sampling rate of all the speech data was 16 kHz.

3. Model

We briefly explain four types of CNN structures: F-CNN, T-CNN, TF-CNN, and PTF-CNN. The CNN has one convolutional layer, one max pooling layer, and one fully connected layer before the softmax layer. For CLSTM-RNN, two LSTM layers were used on top of one convolutional and one max pooling layers instead of the fully connected layer. For all neural networks, the input is 40 log mel filterbank energy and their first and second derivatives with 9 context window.

3.1. Convolutional neural network (CNN)

Frequency-domain CNN (F-CNN) applies convolution and pooling along the frequency axis, and therefore, it can extract useful spectral features while reducing frequency variance. We used a 8×1 frequency filter, non-overlapping max pooling with a pooling size of 3, and 90 feature maps.

Time-domain CNN (T-CNN) applies convolution and pooling along the time axis, and therefore, it can represent modulating characteristics while keeping invariance to a small shift in time. We used a 1×4 time filter, non-overlapping max pooling with a size of 3, and 30 feature maps.

Time-frequency CNN (TF-CNN) applies convolution and pooling along both the time-frequency axis, and therefore, it can extract robust features that are insensitive to a small shift in the time-frequency axis. We used a 8×4 time-frequency filter, non-overlapping max pooling with a size of 3×3, and 40 feature maps.

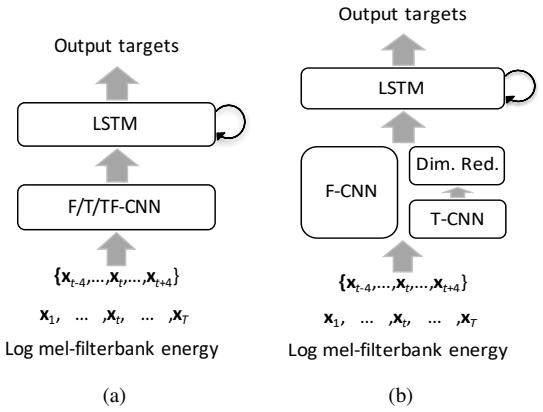

Figure 1: *CLSTM structure for (a) F/T/TF-CLSTM and (b) PTF-CLSTM.*

Parallel TF-CNN (PTF-CNN) has two CNNs that are F-CNN and T-CNN [12]. The units of the max pooling layer in F-CNN and T-CNN are linked with the fully connected layer. In general, the number of units of the max pooling layer in T-CNN are much larger than with F-CNN. We added a linear layer to reduce feature dimension, before passing this to the fully connected layer. We set the output size of the linear layer as the same dimension with F-CNN. F-CNN and T-CNN have the same structure with above mentioned networks.

3.2. Convolutional LSTM-RNN (CLSTM-RNN)

CLSTM-RNN uses 2 LSTM-RNN layers to summarize temporal patterns on top of the CNNs instead of a fully connected layer. We combined the four types of CNNs with LSTM-RNN, resulting in F-CLSTM-RNN, T-CLSTM-RNN, TF-CLSTM-RNN, and PTF-CLSTM-RNN, respectively. The schematic diagram of this structure are represented in Figure 1.

CLSTM-RNN is able to capture the key characteristics of dysarthric speech for speech recognition by modeling long-range temporal structures with time-frequency shift-robust features. Therefore, CLSTM-RNN may be effective in recognizing dysarthric speech.

3.3. Experimental setup

We used HMM-based dysarthric speech recognition systems where each state can be modeled by GMM or neural networks. We compared four types of ASR systems: GMM-HMM, DNN-HMM, CNN-HMM, and CLSTM-RNN-HMM. It consists of 719 tied-state (senone) left-to-right triphone HMMs, where each HMM has 3 states. The senones were obtained using the decision tree-based state tying method. GMM-HMM was trained using 39 dimensional mel-frequency cepstral coefficients, consisting of 12 cepstral coefficients, 1 energy term, and their first and second derivatives with frame size of 25 milliseconds and shift size of 10 milliseconds. DNN-HMM was trained using 40 dimensional log mel-filterbank energy features and their first and second derivatives with a context window of 9 frames. The DNN had 3 hidden layers with 512 hidden units at each layer and the 719 dimensional softmax output layer, corresponding to the senones of the GMM-HMM system. In preliminary experiments, we tested from 1 to 6 layers with 256, 512, and 1,024 hidden units at each layer and obtained the best re-

Table 1: *PERs (%) with training set combination on GMM*

Training data	Speech intelligibility			SA	GA
	High	Mid	Low		
Normal	68.9	72.1	79.1	71.3	73.4
Dysarthric	50.4	53.2	72.0	54.6	58.5
Mixed	**42.0**	**45.8**	**67.6**	**47.1**	**51.8**

Table 2: *PERs (%) on CNN*

Model	Speech intelligibility			SA	GA
	High	Mid	Low		
GMM	42.0	45.8	**67.6**	47.1	51.8
DNN	35.9	41.9	71.4	43.1	49.7
F-CNN	35.4	41.1	70.7	42.5	49.0
T-CNN	34.9	42.2	71.2	42.6	49.4
TF-CNN	33.5	**40.9**	71.4	**41.4**	**48.6**
PTF-CNN	**33.4**	41.5	71.9	41.6	48.9

sults on the 3 hidden layers with 512 hidden units. The parametric rectified linear unit (PReLU) activation function was used and the network was trained using backpropagation.

For each CNN, a variety of filter sizes and feature maps were tested and we set the parameters as in Section 3.1. One fully connected layer with 512 hidden units was used on top of the CNN. LSTM had 2 hidden layers with 320 LSTM cells plus 200 recurrent projection units [14] at each layer and the 719 dimensional softmax output layer. The parameters were trained using backpropagation through time. For CLSTM-RNN, we used the same structure of CNN and replaced the fully connected layer with 2 LSTM-RNN layers. The bigram phoneme language model was used for the phoneme sequence recognition. The bigram language model was trained using the TIMIT training set. The training and decoding were performed using the Kaldi speech recognition toolkit [21].

Phoneme error rates (PERs) were used as the performance measure of dysarthric speech recognition. Leave-one-subject-out cross validation was used to perform speaker-independent phoneme recognition in the experiment. We excluded the session data with low speech intelligibility during training because adding these data to the training set degraded the performance. Thus, we only used the session data with high and mid speech intelligibility from each speaker as a training set. The average performance of cross validations was reported as the overall performance.

4. Results and Discussion

4.1. Baseline

We first explored the effect of a training set to construct a better baseline. We compared normal training data from 7 normal speakers, dysarthric training data from 8 dysarthric speakers with high and mid speech intelligibility sessions (cross validation training set), and mixed training data (normal + dysarthric) on the GMM-based ASR system in Table 1. We averaged test sessions' performance depending on their speech intelligibility level. Here, SA indicates the average of all the test sessions and GA means the average of three speech intelligibility groups. Because the number of sessions associated with each speech intelligibility group is different, GA can show balanced error rates over the groups. The lowest PER in each column is in bold. As shown in Table 1, mixed training produces better results over the normal and dysarthric training conditions. The small number of training data for dysarthric speech is usually available, so using a larger number of normal/out-of-domain speech data is helpful to train acoustic models [5, 22, 23]. In the following experiments, mixed training is used as a default setting.

Table 3: *PERs (%) on CLSTM-RNN*

Model	Speech intelligibility			SA	GA
	High	Mid	Low		
LSTM	29.5	33.8	63.4	36.0	42.2
F-CLSTM	**25.2**	**25.5**	60.4	31.1	37.0
T-CLSTM	30.6	37.9	59.4	37.0	42.6
TF-CLSTM	25.8	26.3	**54.1**	**30.6**	**35.4**
PTF-CLSTM	27.0	26.6	61.9	32.7	38.5

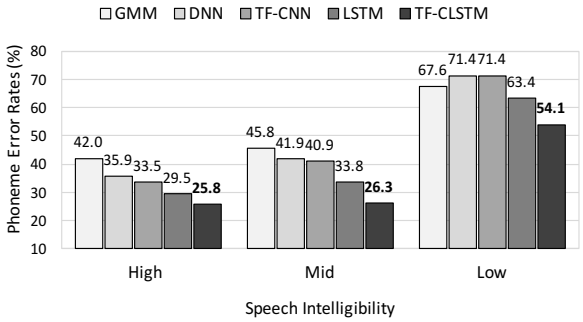

Figure 2: *PERs on the selected models.*

We also measured session-dependent performance for dysarthric speakers using only each session data based on GMM-HMM. The performance was evaluated based on leave-two-utterances-out cross validation for each session. The average of all the test sessions (SA) was 51.0% and the average of three speech intelligibility groups (GA) was 52.0%, which show worse performance than mixed training condition. In addition, the PER of normal speakers was 44.4%, which was obtained using only normal speech data through leave-one-subject-out cross validation (7 cross validation).

4.2. Effect of CNN

Table 2 compares the performance of ASR systems based on the four types of CNNs and DNN. As can be seen, all the CNNs outperformed DNN on both SA and GA. Specifically, F-CNN was slightly better than T-CNN. Considering the frequency and time convolution together (i.e., TF/PTF-CNN) produced better performance than each one. TF-CNN was the best on both SA and GA, producing PERs of 41.4% and 48.6%, respectively. Interestingly, GMM was better than deep models for the low speech intelligibility group while deep models were much better than GMM for the high and mid speech intelligibility groups. This indicates that DNN- and CNN-based acoustic models help to better discriminate between phonemes for the high and mid speech intelligibility groups. However, these types of deep models are still challenging in modeling severely unintelligible speech.

4.3. Effect of CLSTM-RNN

Table 3 shows the performance of LSTM-RNN and CLSTM-RNN. As can be seen, almost all the CLSTM-RNN models were better than LSTM-RNN. T-CLSTM-RNN produced better results than LSTM-RNN for the low intelligible speech whereas for the high and mid intelligible speech, its performance

Table 4: *PERs (%) of the selected models for each individual session*

Spk	Sess.	Intell. (%)	PER (%)		
			CNN	LSTM	CLSTM
SPK1	S1	95.4 (H)	25.8	19.1	**16.7**
SPK2	S1	80.0 (M)	43.8	40.0	**26.9**
SPK3	S1	100 (H)	**44.1**	51.2	51.8
	S2	100 (H)	**26.3**	31.5	28.7
	S3	100 (H)	45.5	48.0	**36.8**
SPK4	S1	98.1 (H)	31.2	31.3	**27.3**
	S2	97.2 (H)	13.7	11.3	**10.5**
	S3	79.0 (M)	24.7	22.9	**16.7**
SPK5	S1	99.0 (H)	31.2	20.1	**19.4**
	S2	98.1 (H)	42.9	25.5	**14.1**
	S3	14.5 (L)	66.7	56.5	**29.5**
	S4	0 (L)	78.8	74.7	**74.6**
SPK6	S1	94.5 (H)	33.5	20.3	**16.4**
	S2	80.9 (M)	43.5	28.9	**26.8**
	S3	23.6 (L)	68.9	59.2	**58.4**
SPK7	S1	99.0 (H)	43.6	35.5	**27.9**
SPK8	S1	96.3 (H)	30.7	**29.1**	34.2
SPK9	S1	79.0 (M)	51.8	43.7	**35.1**
Average			41.4	36.0	30.6
Standard Deviation			16.8	16.6	16.5

was worse than LSTM-RNN. When we consider frequency-domain convolution (i.e., F-CLSTM-RNN, TF-CLSTM-RNN, and PTF-CLSTM-RNN), the performance was much improved. Specifically, F-CLSTM-RNN was the best on the high and mid intelligible speech while TF-CLSTM-RNN was the best on the low intelligible speech. In addition, we obtained the lowest PER on TF-CLSTM-RNN in terms of SA and GA, showing 30.6% and 35.4%, respectively.

Figure 2 summarizes the PERs on the selected models (GMM, DNN, TF-CNN, LSTM-RNN, and TF-CLSTM-RNN). As shown in Figure 2, LSTM-RNN provided better results than GMM, DNN, and TF-CNN for all the speech intelligibility groups. This implies that modeling temporal structures by LSTM-RNN may be more important in recognizing dysarthric speech. When we used TF-CLSTM-RNN, we were able to obtain the best performance for all the groups, producing 12.5%, 22.1%, and 14.6% relative improvements in the PER over LSTM-RNN for high, mid, and low intelligible speech, respectively. This indicates TF-CLSTM-RNN can effectively capture the time-frequency characteristics over time even for highly unintelligible speech.

4.4. Evaluation of each individual session

PERs of the selected models (TF-CNN, LSTM-RNN, and TF-CLSTM-RNN) for each individual are presented in Table 4. Here, "Spk" means speaker ID and "Sess." indicates their session ID. The average duration between successive sessions is 6 months. "Intell." is the perceptual speech intelligibility score in percent and the letters in parenthesis indicate their speech intelligibility groups (i.e., H: high, M: mid, and L: low). The lowest PER in each row is in bold. As can be seen, TF-CLSTM-RNN gave the best results for almost all the speakers/sessions including low intelligible speech sessions. In particular, TF-CLSTM-RNN produced a 47.7% relative improvement in the PER over LSTM-RNN for the 3rd session data of SPK5 (i.e., SPK5-S3, speech intelligibility of 14.5%). However, it was still very challenging for extremely low intelligible speech (SPK5-S4, speech intelligibility of 0%).

4.5. Discussion

We observed high variability in speech intelligibility and ASR performance across sessions within individual speakers because the rate of disease progression varies among speakers with ALS. For example, SPK5's speech intelligibility declined from 99.0% to 0% and their PER varied from 14.1% to 74.6%. As the disease progresses, tongue body and jaw movement patterns of ALS patients become different from the articulatory motion patterns of normal speakers [24]. For this reason, the acoustic characteristics in early sessions from speakers (e.g., SPK5-S1 and SPK5-S2) are similar to those of normal speakers. Contrastingly, the acoustic characteristics in late sessions from speakers (e.g., SPK5-S3 and SPK5-S4) show different patterns compared with early sessions: poor articulation, long pause between words, and low speaking rates (about 50% lower) [24]. All of which lower ASR performance and the capability to use ASR as ALS progresses.

This current study was conducted in the context of a speaker-independent ASR task to evaluate the generality of the recognition models. Our experimental results demonstrated that CLSTM-RNN has the potential to improve the ASR performance as a speaker-independent acoustic model for the patients with ALS. To further improve the ASR accuracies, techniques for session/speaker variability compensation including acoustic feature transformation [25, 26], acoustic model adaptation [27], and pronunciation variation modeling [27,28] can be further applied. We speculate that the results may improve once a larger training dataset from more ALS patients is obtained. A further study with larger data size and more patients with diverse severity of dysarthria is needed to verify this finding.

5. Conclusions and Future Work

In this paper, we investigated the effectiveness of CLSTM-RNN for dysarthric speech recognition. We considered four types of CLSTM-RNN, including F-CLSTM-RNN, T-LSTM-RNN, TF-LSTM-RNN, and PTF-LSTM-RNN. A series of experiments was performed in terms of the PER on 18 sessions speech data from 9 ALS patients. Experimental results showed that the CLSTM-RNN provides meaningful improvement over both the CNN alone and the LSTM-RNN alone. We achieved the best overall performance on TF-CLSTM-RNN (PERs of 30.6% and 35.4% for SA and GA, respectively). Our approach presents a possibility in effectively modeling dysarthric speech (even low intelligible speech) in a speaker-independent way. Future directions include 1) a test of the CLSTM-RNN approach using a larger dataset collected from more subjects, 2) applying speaker adaptation/normalization techniques [27], and 3) using articulatory information [25, 29].

6. Acknowledgements

This work was supported by the National Institutes of Health through grants R03 DC013990 and R01 DC013547 and American Speech-Language-Hearing Foundation through a New Century Scholar Research Grant. We would like to thank Dr. Jordan R. Green, Dr. Thomas F. Campbell, and Dr. Yana Yunusova, Kristin Teplansky, Jennifer McGlothlin and the volunteering participants.

7. References

[1] J. R. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. Elsevier Health Sciences, 2013.

[2] M. Hasegawa-Johnson, J. Gunderson, A. Perlman, and T. Huang, "HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 3, 2006, pp. 1060–1063.

[3] M. J. Kim, J. Yoo, and H. Kim, "Dysarthric speech recognition using dysarthria-severity-dependent and speaker-adaptive models." in *Interspeech*, 2013, pp. 3622–3626.

[4] F. Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 947–960, 2011.

[5] M. Kim, J. Wang, and H. Kim, "Dysarthric speech recognition using kullback-leibler divergence-based hidden markov model," in *Proc. of Interspeech*, 2016, pp. 2671–2675.

[6] E. Yılmaz, M. Ganzeboom, C. Cucchiarini, and H. Strik, "Multi-stage dnn training for automatic recognition of dysarthric speech," *Proc. Interspeech*, pp. 2685–2689, 2017.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[9] T. N. Sainath, B. Kingsbury, A.-r. Mohamed, G. E. Dahl, G. Saon, H. Soltau, T. Beran, A. Y. Aravkin, and B. Ramabhadran, "Improvements to deep convolutional neural networks for LVCSR," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 315–320.

[10] L. Tóth, "Combining time-and frequency-domain convolution in convolutional neural network-based phone recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 190–194.

[11] T. Nakashika, T. Yoshioka, T. Takiguchi, Y. Ariki, S. Duffner, and C. Garcia, "Dysarthric speech recognition using a convolutive bottleneck network," in *12th IEEE International Conference on Signal Processing (ICSP)*, 2014, pp. 505–509.

[12] V. Mitra and H. Franco, "Time-frequency convolutional networks for robust speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015, pp. 317–323.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014, pp. 14–18.

[15] C. Espana-Bonet and J. A. Fonollosa, "Automatic speech recognition with deep neural networks for impaired speech," in *Third International Conference on Advances in Speech and Language Technologies for Iberian Languages*. Springer, 2016, pp. 97–107.

[16] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4580–4584.

[17] K. J. Han, S. Hahm, B.-H. Kim, J. Kim, and I. Lane, "Deep learning-based telephony speech recognition in the wild," in *Proc. Interspeech*, 2017, pp. 1324–1327.

[18] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 2392–2396.

[19] J. R. Green, Y. Yunusova, M. S. Kuruvilla, J. Wang, G. L. Pattee, L. Synhorst, L. Zinman, and J. D. Berry, "Bulbar and speech motor assessment in ALS: Challenges and future directions," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 14, no. 7-8, pp. 494–500, 2013.

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

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Waikoloa, USA, 2011, pp. 1–4.

[22] H. Christensen, M. Aniol, P. Bell, P. D. Green, T. Hain, S. King, and P. Swietojanski, "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech." in *Proc. Interspeech*, 2013, pp. 3642–3645.

[23] E. Yilmaz, M. Ganzeboom, C. Cucchiarini, and H. Strik, "Combining non-pathological data of different language varieties to improve dnn-hmm performance on pathological speech," in *Proc. Interspeech*, 2016, pp. 218–222.

[24] J. Lee, M. Bell, and Z. Simmons, "Articulatory kinematic characteristics across the dysarthria severity spectrum in individuals with amyotrophic lateral sclerosis," *American Journal of Speech-Language Pathology*, vol. 27, pp. 258–269, 2018.

[25] S. Hahm, H. Daragh, and J. Wang, "Recognizing dysarthric speech due to amyotrophic lateral sclerosis with across-speaker articulatory normalization," in *Proc. the ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies*, 2015, pp. 47–54.

[26] B. Vachhani, C. Bhat, B. Das, and S. K. Kopparapu, "Deep autoencoder based speech features for improved dysarthric speech recognition," *Proc. Interspeech*, pp. 1854–1858, 2017.

[27] M. Kim, Y. Kim, J. Yoo, J. Wang, and H. Kim, "Regularized speaker adaptation of KL-HMM for dysarthric speech recognition," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 9, pp. 1581–1591, 2017.

[28] S. O. C. Morales and S. J. Cox, "Modelling errors in automatic speech recognition for dysarthric speakers," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, pp. 308–340, 2009.

[29] M. Kim, B. Cao, and J. Wang, "Multi-view representation learning via canonical correlation analysis for dysarthric speech recognition." in *Proc. International Conference on Mechatronics and Intelligent Robotics*, 2018.