INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

On Improving Acoustic Models For TORGO Dysarthric Speech Database

Neethu Mariam Joy, S. Umesh, Basil Abraham

Indian Institute of Technology-Madras, India
{ee11d009,umeshs,ee11d032}@ee.iitm.ac.in

Abstract

Assistive technologies based on speech have been shown to improve the quality of life of people affected with dysarthria, a motor speech disorder. Multiple ways to improve Gaussian mixture model-hidden Markov model (GMM-HMM) and deep neural network (DNN) based automatic speech recognition (ASR) systems for TORGO database for dysarthric speech are explored in this paper. Past attempts in developing ASR systems for TORGO database were limited to training just monophone models and doing speaker adaptation over them. Although a recent work attempted training triphone and neural network models, parameters like the number of context dependent states, dimensionality of the principal component features etc were not properly tuned. This paper develops speaker-specific ASR models for each dysarthric speaker in TORGO database by tuning parameters of GMM-HMM model, number of layers and hidden nodes in DNN. Employing dropout scheme and sequence discriminative training in DNN also gave significant gains. Speaker adapted features like feature-space maximum likelihood linear regression (FMLLR) are used to pass the speaker information to DNNs. To the best of our knowledge, this paper presents the best recognition accuracies for TORGO database till date.

Index Terms: Dysarthria, TORGO, GMM-HMM, DNN

1. Introduction

Dysarthria is a speech disorder caused by degeneration in neuromuscular control of motor speech articulators. This causes difficulty in controlling muscles, such as tongue and lips, involved in speech production [1]. The produced speech is characterized by irregular articulation of phonemes [2][3], monotone nature or excessive variation of loudness and pitch [4][5], slurring, mumbling and slow speaking rate. The speech will be dysfluent due to prolongations of words, stop-gaps and repetitions. Listeners who are not familiar with the particular impairments of the speaker will find it hard to recognize what is being said. Often the disabled person is also physically incapacitated and has limited hand movement and co-ordination capability.

Speech based assistive technologies can be very helpful in this scenario [6][7]. Using automatic speech recognition (ASR) systems trained on normal un-impaired speech [7] will not be useful as dysarthric speech and normal speech are different [8][9][10]. When a normal speech ASR system is used for dysarthric speech, the word error rates (WER) are $26.2\% - 81.8\%$ higher than the normal speech WER [11]. Physical fatigue, frustration of speaker, and severity level of the disability increases the variabilities in dysarthric speech and poses a challenge in building accurate speech based assistive technologies for dysarthric speakers. The trade-off would be to build such systems tailored to the specifics of the speaker's impairment which leverages knowledge from normal speech based technologies as much as possible [3][8][12][13][14][15].

Recent research have shown that dysarthric assistive technologies with neural networks as the mainframe gives improved performance. In dysarthric speech recognition, neural networks tackle problems of complex acoustic modeling, extracting discriminative features, correcting pronunciation errors and transferring knowledge from other standard speech corpus. Bottleneck features extracted from convolutional neural network (CNN) are used for acoustic modeling in [16] to reduce the articulatory errors arising from strain on the speech muscles in dysarthric speakers. In [17], a multi-views multi-learners approach using multi-nets artificial neural networks (ANN) was proposed, which employs several ANN as learners to deal with the complexity of dysarthric speech. Christensen *et al.* [18] included features generated by deep neural network (DNN) trained on out-of-domain speech data to improve the dysarthric ASR trained on Universal Access (UA) speech database [19].

In this paper, we aim to develop improved GMM-HMM and DNN-HMM acoustic models for TORGO dysarthric speech database [20]. The first attempt in developing a ASR system for TORGO was reported in Mengistu *et al.* [21] where a monophone GMM-HMM model was trained and was then maximum likelihood linear regression (MLLR) adapted to dysarthric speakers. Later, in España-Bonet *et al.* [22], neural network architectures like DNN, CNN, time-delay DNN (TDNN), and long short term memory (LSTM) recurrent neural network (RNN) were used in developing a dysarthric ASR system for TORGO database.

Our paper presents experiments performed on similar lines to España-Bonet *et al.* [22], but with significant reduction of 17.62% relative in WER by virtue of tuning parameters of GMM-HMM and DNN-HMM. We varied the frame rate during feature extraction, chose word position independent phones for acoustic modeling, varied the number of context dependent states and dimensionality of principal component features. The afore mentioned parameters were optimized for each dysarthric speaker-dependent GMM-HMM model in TORGO database and was later used in DNN-HMM modeling. The number of hidden layers and nodes were optimized and dropout and sequence discrimation training strategies were adopted to improve DNN-HMM modeling.

The organization of the paper is as follows. Section 2 describes the salient features of the TORGO database. The details of various factors involved in tuning the GMM-HMM and DNN-HMM are given in section 3 and 4 respectively. Section 5 summarizes the major contributions of the paper.

2. TORGO Dysarthric Speech Database

The TORGO database [20] consists of aligned acoustic and articulatory recordings from 15 speakers of which eight (5 males, 3 females) are dysarthric and the remaining seven (4 males, 3 females) are control speakers without any disorder [23]. The degree of severity of the disorder for each of the dysarthric speakers were evaluated by a speech-language pathologist in-terms of

http://dx.doi.org/10.21437/Interspeech.2017-878

Table 1: *Details of TORGO dysarthric speech database*

Speaker	Dysarthric Speakers								Control Speakers						
	F01	**M01**	**M02**	**M04**	**M05**	**F03**	**F04**	**M03**	**FC01**	**FC02**	**FC03**	**MC01**	**MC02**	**MC03**	**MC04**
Disorder	Severe	Severe	Severe	Severe	Severe-Moderate	Moderate	Mild	Mild	None	None	None	None	None	None	None
#Utterance	228	739	772	659	610	1097	675	806	296	2183	1924	2141	1112	1661	1614

*F: female speaker, M: male speaker, FC: female control speaker, MC: male control speaker

the overall clinical intelligibility and the motor functions of the articulators as per the Frenchay dysarthria assessment [24]. The results of this assessment are as shown in Table 1.

During multiple sessions, at-least three hours of speech was recorded from each speaker in-terms of single words, sentences and their description of contents in a photograph. On an average, 415 and 800 utterances were recorded from each dysarthric and control speaker respectively. The precise number of audio files recorded per speaker is given in Table 1. The single words considered were English digits, international radio alphabets, twenty most frequent words in British National Corpus (BNC), and a set of words selected by Kent *et al.* [25] to account for relevant phonetic contrasts. The sentences were taken from Yorkston-Beukelman assessment of intelligibility [26] and the TIMIT database [27]. The subjects were also asked to describe the contents of a few photographs in his/her own words, to include dictation style speech into the database.

3. GMM-HMM Acoustic Modeling

For each of the eight dysarthric speakers, a speaker dependent GMM-HMM model was built using Kaldi toolkit [28] using the recipe[1] mentioned in [22] . In each case, all the utterances from the particular dysarthric speaker under consideration was used for decoding purposes. The audio files from the remaining seven dysarthric and seven control speakers were used for training. For example, while building F01 dysarthric speaker's GMM-HMM model, 228 utterances from F01 was used for decoding and the remaining 16244 utterances from all the other speakers were used for training the acoustic model. In the following subsections, we describe in detail the various parameters that were tuned to achieve improvements in recognition accuracy of the GMM-HMM model.

3.1. Frame Rate

For each frame, a frame length and frame shift of 25ms and 10ms respectively were used for extracting 13-dimensional Mel frequency cepstral coefficients (MFCC) for control speakers. España-Bonet *et al.* [22] proposed that a frame length and frame shift of 25ms and 15ms respectively need to be used for extracting MFCC features for dysarthric speakers. This claim was based on the fact that dysarthric speech is characterized by slow speaking rate and hence widening the frame shift can homogenize the differences between dysarthric and control speakers.

In Table 2, we verify this claim by generating two sets of features for dysarthric speakers, one with 10ms frame shift and the other with 15ms frame shift. This analysis was done for F01 dysarthric speaker's GMM-HMM model which was trained with 1800 context dependent states and a total of 9000 Gaussians as mentioned in España-Bonet *et al.* [22]. While GMM-HMM model in row-5 of Table 2 seems to show significant gains over other models, we infer that this is not by virtue of just increasing the frame shift as claimed in [22].

[1] https://github.com/cristinae/ASRdys

Table 2: *Varying frame shift of dysarthric speech and choosing between position dependent and independent phones for monophone and triphone models of F01 severe dysarthric speaker*

Phone	Frame Shift		%WER	
	Dysarthric	**Control**	**Monophone**	**Triphone**
España-Bonet et al. [22]				
Pos-Dep	15ms	10ms	70.86	70.68
Varying frame shift and choosing position dependency of phones				
Pos-Dep	15ms	10ms	67.45	64.39
Pos-Dep	10ms	10ms	63.67	68.71
Pos-Ind	10ms	10ms	62.41	62.95
Pos-Ind	15ms	10ms	**51.80**	**45.32**
Relative Reduction in WER wrt [22]			26.89	35.88

*All triphone models have 1800 tied states and 9000 Gaussians.

3.2. Word Position Dependency of Phones

España-Bonet *et al.* [22] uses word position dependent phones for acoustic modeling. This practice is generally adopted in Kaldi toolkit to expand the phone list by including various versions of a phone depending on the position of that phone within a word. For instance, for a phone aa, this extended phone list will include aa_B, aa_E, aa_I and aa_S, indicating whether the phone occurs in the beginning, end, internal of a word or as a singleton phone respectively.

Due to the slow speaking rate, slurring and stop-gaps in dysarthric speech, the co-articulation is lower and often phones occur in isolation when compared to control speech. Hence, modeling for word-position dependency of phones is an overkill. To verify our claim, the analysis in section 3.1 was repeated by considering position dependent and independent phones. GMM-HMM model in row-5 of Table 2 which uses position independent phones and a frame shift of 15ms for dysarthric speakers gave relative WER reduction of 35.88% compared to [22]. Hence in our experiments, we have used frame shift of 15ms for dysarthric speakers and position independent phone sets for acoustic modeling.

3.3. Number of Context Dependent States

In [22], España-Bonet *et al.* has fixed the number of context dependent states (tied states) to 1800 for all the dysarthric speaker acoustic models. However, as the training data characteristics (number of utterances, speakers involved) changes for each dysarthric speaker's GMM-HMM, tuning for the number of context dependent states for each such GMM-HMM model is necessary. As these context dependent states are further carried on for neural network modeling, obtaining a better performing GMM-HMM model is of paramount importance.

Table 3: *Tuning for the number of context dependent states for various dysarthric speaker GMM-HMM models*

Acoustic Model	# Tied States	Dysarthric Test Speakers (%WER)							
		F01	M01	M02	M04	M05	F03	F04	M03
España-Bonet et al. [22]									
Monophone	-	70.86	80.10	76.55	88.62	77.71	57.02	29.10	43.32
Triphone	1800	70.68	91.18	81.09	88.62	84.59	41.80	18.62	26.01
Proposed Tuning of Context Dependent States									
Monophone	-	51.80	70.20	49.95	80.92	66.79	48.43	27.05	20.28
Triphone	400	46.40	65.01	47.84	85.97	**65.37**	35.93	**17.45**	13.06
	500	47.84	**62.79**	**45.21**	83.50	66.34	36.44	17.80	11.97
	600	**45.50**	68.20	49.05	84.23	70.38	35.22	17.45	12.76
	700	46.22	69.23	48.84	**79.83**	77.26	34.79	17.92	12.91
	800	49.82	67.23	48.47	83.56	71.88	**33.68**	17.97	**11.52**
	900	45.68	71.61	47.79	85.13	83.25	35.18	18.09	11.72
	1800	45.32	75.77	53.32	80.98	85.19	34.61	18.44	12.02
Relative Reduction in WER wrt [22]		35.63	31.14	44.25	9.92	22.72	19.43	6.28	55.71

Table 4: *Tuning for the dimensionality of LDA features for various dysarthric speaker GMM-HMM models*

Acoustic Model	LDA Dimension	Dysarthric Test Speakers (% WER)							
		F01	M01	M02	M04	M05	F03	F04	M03
España-Bonet et al. [22]									
Triphone	-	70.68	91.18	81.09	88.62	84.59	41.80	18.62	26.01
LDA	40	76.80	79.12	83.67	88.68	96.71	53.08	18.97	32.59
Proposed Tuning of LDA Feature Dimension									
Triphone	-	45.50	62.79	45.21	79.83	65.37	33.68	17.45	11.52
LDA	40	55.22	68.90	53.79	85.37	76.07	38.69	18.03	11.57
	35	50.36	70.15	50.32	**82.24**	75.54	36.51	**17.04**	12.07
	30	**50.00**	66.36	47.73	82.78	68.74	36.47	18.85	**11.42**
	25	50.54	**61.93**	**47.68**	82.90	**63.87**	**36.15**	18.33	11.77
Relative Reduction in WER wrt [22]		34.89	21.73	43.01	7.26	33.95	31.89	10.17	64.96

In Table 3, the number of context dependent states in the triphone GMM-HMM model was varied from 400 to 1800. The total number of Gaussians was fixed at 9000. Tied states capture the co-articulation effects in speech. As dysarthric speech has less co-articulation compared to normal speech, using lower number of tied states gives better performance as shown in Table 3. We observe that severe dysarthric speakers uses less number of tied states as their speech has low degree of co-articulation with phones often occurring in isolation. Tuning the number of tied states has significantly decreased the WER of triphone models when compared to [22].

3.4. Dimension of LDA Features

Once 13-dimensional MFCC features with suitable frame shift was extracted as described in section 3.1, we augment it's first and second order derivatives to obtain a 39-dimensional vector, which is then mean normalized at speaker level. Consecutive nine frames of this feature is then spliced together. It is then projected down to a M-dimensional vector using linear discriminant analysis (LDA) and further diagonalized by maximum likelihood linear transformation (MLLT). These M-dimensional features are called LDA features.

In [22], M was fixed at 40. Table 4 shows the effect of varying M from 25 to 40 in steps of 5 for various dysarthric speaker GMM-HMM models trained on LDA features. It can be seen that reducing LDA dimension in the range of 30 to 25 shows improvement over 40-dimensional LDA features. This can be attributed to the constrained nature of data in TORGO database which limits the number of principal component dimensions. In our experiments, we have chosen the dimensionality of LDA features as 30.

3.5. Speaker Normalized FMLLR Features

Table 5: *WER for various dysarthric speaker GMM-HMMs*

Feature	Dysarthric Test Speakers *(% WER)*							
	F01	M01	M02	M04	M05	F03	F04	M03
España-Bonet et al. [22]								
MFCC	70.68	91.18	81.09	88.62	84.59	41.80	18.62	26.01
LDA	76.80	79.12	83.67	88.68	96.71	53.08	18.97	32.59
FMLLR	47.30	78.91	68.49	81.16	97.16	42.88	**13.29**	17.06
Proposed GMM-HMM Model After Parameter Tuning								
MFCC	45.50	62.79	45.21	79.83	65.37	33.68	17.45	11.52
LDA	50.00	66.36	47.73	82.78	68.74	36.47	18.85	11.42
FMLLR	**29.68**	**62.03**	**42.31**	**74.17**	**73.75**	**32.86**	14.58	**7.86**
Relative Reduction in WER wrt [22]								
FMLLR	37.25	21.39	38.22	8.61	24.09	23.37	-9.71	53.93

The LDA features extracted as mentioned in section 3.4 are further speaker normalized via feature-space maximum likelihood linear regression (FMLLR). This removes the speaker variabilities in the features. As we have considered 30-dimensional

LDA features in our experiments, the FMLLR features will also be 30-dimensional. In [22], 40-dimensional FMLLR features were used. Table 5 compares the WER of various GMM-HMM models trained on MFCC, LDA and FMLLR features for different dysarthric test speakers. GMM-HMMs trained on FMLLR features are superior to other models as they effectively remove speaker-dependent variabilities and train a compact model.

4. DNN-HMM Acoustic Modeling

The 30-dimensional FMLLR features extracted were used to train DNN models for various dysarthric speakers. The number of units in the output softmax layer equals the the number of tied states in phonetic decision tree as mentioned in section 3.3. The frame-level alignment information was obtained from the GMM-HMM model trained on FMLLR features (Table 5).

The entire training data was initially randomized at frame level. The DNN parameters are initialized by layer-wise RBM pretraining. Supervised training uses stochastic gradient descent with a mini-batch size of 256 frames and learning rate of 0.008. The FMLLR features were stacked over a context window of 11 frames (± 5) and are fed as input to the DNN. Weighted finite state transducer (WFST) based graph generated for GMM-HMM model was used for decoding the DNN models by scaling DNN posteriors with class priors computed from alignments. In the following subsections, we describe the various factors which contributed to improvement in DNN-HMM recognition performance.

4.1. Number of Hidden Layers and Nodes

In [22], the DNN model was trained with 6 hidden layers and 1024 neurons per layer, with the number of tied states fixed at 1800. In section 3.3, we showed that the number of tied states has to be tuned for each dysarthric speaker GMM-HMM model. We also varied the complexity of the DNN model by reducing the number of hidden layers and nodes per layer as shown in Table 6. It can be seen that, a less complex model than the one in [22] gave 27.72% relative reduction in WER. In our experiments, we used 4 hidden layers with sigmoid activation functions and 1024 neurons in each layer for all DNNs.

Table 6: *Varying the number of hidden layers and neurons per layer of DNN-HMM models trained on FMLLR features for F01 severe dysarthric speaker*

#Layers	#Nodes	#Tied States	%WER
España-Bonet et al. [22]			
6	1024	1800	39.57
Proposed Tuning for number of hidden layers and nodes			
3	512	600	29.32
4	512	600	29.68
4	1024	600	**28.60**
4	2048	600	28.96
5	1024	600	30.22
5	2048	600	31.12
Relative Reduction in WER wrt [22]			27.72

4.2. Dropout for DNN Generalization

Dropout technique randomly omits a certain percentage of the neurons in each hidden layer of a DNN during training. This reduces the dependency of each neuron on other neurons to detect patterns. Applying dropout was shown to increase the generalizing ability of DNNs. We apply a dropout factor of 0.2 for the first four DNN training epochs. On an average, the WER drops by 0.54% absolute when dropout is applied as shown in Table 7. This is because the test dysarthric speaker is unknown for the train data and hence the DNN model must be capable of generalizing the information learned from other dysarthric speakers to avoid overfitting the DNN model towards the train data.

Table 7: *WER for various dysarthric speaker DNN-HMMs*

Dropout	sMBR	Dysarthric Test Speakers (% WER)							
		F01	M01	M02	M04	M05	F03	F04	M03
España-Bonet et al. [22]									
No	No	39.57	62.20	42.89	69.05	62.60	39.30	13.06	17.71
No	Yes	35.61	62.30	47.95	69.30	**62.53**	37.01	10.95	12.76
Proposed DNN-HMM Model									
No	No	28.60	49.81	39.78	74.89	71.73	34.43	12.70	7.72
No	Yes	**25.54**	**45.70**	**36.72**	**67.85**	66.19	**31.07**	**10.83**	**6.48**
Yes	No	27.34	48.89	39.62	74.89	69.78	34.43	12.53	7.89
Yes	Yes	25.90	47.49	37.36	69.72	67.17	31.68	10.89	6.48
Relative Reduction in WER wrt [22]									
No	Yes	28.28	26.65	23.42	2.09	-5.85	16.05	1.10	49.22

4.3. Sequence Discriminative Training

Sequence discriminative training can be done on top of DNNs trained using cross-entropy criterion to minimize the errors on state labels in a sequence. We perform 6 iterations of state-level minimum Bayes risk (sMBR) on top of a trained DNN for sequence discrimination. España-Bonet *et al.* [22] has shown improvements with sequence discrimination training, but without dropout applied. From Table 7, it can be seen that sequence discriminative training improves the recognition accuracy of DNNs, with or without the application of dropout.

Comparing with the sMBR applied DNN in [22], the proposed DNN models on an average shows 6.004% absolute and 17.62% relative reduction in WER. The DNN models of all the severe dysarthric speakers in our experiments gave relative reduction of 20.11% in WER on an average compared to [22].

5. Conclusion

In this paper, we have developed GMM-HMM and DNN-HMM acoustic models for ASR based assistive technologies for dysarthric speakers in TORGO database. Compared to the previous attempts of developing such ASR systems in TORGO, the various acoustic models trained here showed significant improvement in recognition accuracies. This is attributed to careful and rigorous tuning of GMM-HMM parameters like frame shift during feature extraction, using word position independent phones, number of context dependent states, dimensionality of LDA features and using speaker normalized FMLLR features for acoustic modeling. DNN models were also tuned for the number of hidden nodes and neurons and was further generalized via dropout and sequence discriminative training. The best tuned DNN models gave relative WER reduction of 17.62% on an average across all dysarthric speaker models compared to the previously published results.

6. References

[1] F. L. Darley, A. E. Aronson, , and J. R. Brown, *Motor Speech Disorders*. W. B. Saunders, 1975.

[2] K. Rosen and S. Yampolsky, "Automatic Speech Recognition and a Review of Its Functioning with Dysarthric Speech," *Augmentative and Alternative Communication*, vol. 16, no. 1, pp. 48–60, 2000.

[3] P. D. Polur and G. E. Miller, "Effect of High-Frequency Spectral Components in Computer Recognition of Dysarthric Speech based on a Mel-Cepstral Stochastic Model," *Journal of Rehabilitation Research and Development*, vol. 42, no. 3, pp. 363–371, 2005.

[4] J. R. Duffy, *Motor Speech Disorders*. Elsevier, 2005.

[5] R. D. Kent, G. Weismer, J. F. Kent, H. K. Voperian, and J. R. Duffy, "Acoustic Studies of Dysarthric Speech: Methods, Progress, and Potential," *Journal of Communication Disorders*, vol. 32, no. 3, pp. 141–186, 1999.

[6] P. C. Doyle, H. A. Leeper, A. L. Kotler, N. ThomasStonell, C. Oneill, M. C. Dylke, and K. Rolls, "Dysarthric Speech: A Comparison of Computerized Speech Recognition and Listener Intelligibility," *Journal of Rehabilitation Research and Development*, vol. 34, no. 3, pp. 309–316, 1997.

[7] L. Ferrier, H. Shane, H. Ballard, T. Carpenter, and A. Benoit, "Dysarthric Speakers Intelligibility and Speech Characteristics in Relation to Computer Speech Recognition," *Augmentative and Alternative Communication*, vol. 11, no. 3, pp. 165–175, 1995.

[8] M. S. Hawley, P. Enderby, P. Green, S. Cunningham, J. C. Simon Brownsell, M. Parker, A. Hatzis, P. O'Neill, and R. Palmer, "A Speech-Controlled Environmental Control System for People with Severe Dysarthia," *Medical Engineering & Physics*, vol. 29, no. 5, pp. 586–593, 2007.

[9] K. Hux, J. Rankin-Erickson, N. Manasse, and E. Lauritzen, "Accuracy of Three Speech Recognition Systems: Case Study of Dysarthric Speech," *Augmentative and Alternative Communication*, vol. 16, no. 3, pp. 186–196, 2000.

[10] E. Sanders, M. B. Ruiter, L. Beijer, and H. Strik, "Automatic Recognition of Dutch Dysarthric Speech: A Pilot Study," in *Proc. INTERSPEECH*, 2002.

[11] F. Rudzicz, "Using Articulatory Likelihoods in the Recognition of Dysarthric Speech," *Speech Communication*, vol. 54, no. 3, pp. 430–444, 2012.

[12] S. Selouani, M. S. Yakoub, and D. D. O'Shaughnessy, "Alternative Speech Communication System for Persons with Severe Speech Disorders," *EURASIP Journal on Advances in Signal Processing*, 2009.

[13] S. O. C. Morales and S. J. Cox, "Modelling Errors in Automatic Speech Recognition for Dysarthric Speakers," *EURASIP Journal on Advances in Signal Processing*, 2009.

[14] P. D. Polur and G. E. Miller, "Investigation of an HMM/ANN Hybrid Structure in Pattern Recognition Application using Cepstral Analysis of Dysarthric (Distorted) Speech Signals," *Medical Engineering & Physics*, vol. 28, no. 8, pp. 741–748, 2006.

[15] M. Hasegawa-Johnson, J. Gunderson, A. Perlman, and T. S. Huang, "HMM-Based and SVM-Based Recognition of the Speech of Talkers With Spastic Dysarthria," in *Proc. ICASSP*, 2006, pp. 1060–1063.

[16] Y. Takashima, T. Nakashika, T. Takiguchi, and Y. Ariki, "Feature Extraction using Pre-Trained Convolutive Bottleneck Nets for Dysarthric Speech Recognition," in *Proc. EUSIPCO*, 2015, pp. 1411–1415.

[17] S. R. Shahamiri and S. S. B. Salim, "A Multi-Views Multi-Learners Approach Towards Dysarthric Speech Recognition Using Multi-Nets Artificial Neural Networks," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 5, pp. 1053–1063, 2014.

[18] H. Christensen, M. B. Aniol, P. Bell, P. Green, T. Hain, S. King, and P. Swietojanski, "Improving Recognition of Disordered Speech with Out-of-Domain Knowledge," in *Proc. INTERSPEECH*, 2013.

[19] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric Speech Database for Universal Access Research," in *Proc. INTERSPEECH*, 2008, pp. 1741–1744.

[20] F. Rudzicz, G. Hirst, P. van Lieshout, G. Penn, F. Shein, A. Namasivayam, and T. Wolff, "TORGO Database of Dysarthric Articulation LDC2012S02," *Linguistic Data Consortium*, 2012.

[21] K. T. Mengistu and F. Rudzicz, "Adapting Acoustic and Lexical Models to Dysarthric Speech," in *Proc. ICASSP*, 2011, pp. 4924–4927.

[22] C. España-Bonet and J. A. R. Fonollosa, "Automatic Speech Recognition with Deep Neural Networks for Impaired Speech," in *Proc. IberSPEECH*, 2016, pp. 97–107.

[23] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO Database of Acoustic And Articulatory Speech from Speakers with Dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[24] P. Enderby, "Frenchay Dysarthria Assessment," *International Journal of Language and Communication Disorders*, vol. 15, no. 3, pp. 165–173, 1980.

[25] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek, "Toward Phonetic Intelligibility Testing in Dysarthria," *Journal of Speech and Hearing Disorders*, vol. 54, pp. 482–499, 1989.

[26] K. M. Yorkston and D. R. Beukelman, *Assesment of Intelligibility of Dysarthric Speech*, 1981.

[27] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1," *Linguistic Data Consortium*, 1993.

[28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. K. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. ASRU*, 2011.