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

Deep Autoencoder based Speech Features for Improved Dysarthric Speech Recognition

Bhavik Vachhani, Chitralekha Bhat, Biswajit Das, Sunil Kumar Kopparapu

TCS Innovation Labs, Mumbai

bhavik.vachhani@tcs.com, bhat.chitralekha@tcs.com, b.das@tcs.com,
sunilkumar.kopparapu@tcs.com

Abstract

Dysarthria is a motor speech disorder, resulting in mumbled, slurred or slow speech that is generally difficult to understand by both humans and machines. Traditional Automatic Speech Recognizers (ASR) perform poorly on dysarthric speech recognition tasks. In this paper, we propose the use of deep autoencoders to enhance the Mel Frequency Cepstral Coefficients (MFCC) based features in order to improve dysarthric speech recognition. Speech from healthy control speakers is used to train an autoencoder which is in turn used to obtain improved feature representation for dysarthric speech. Additionally, we analyze the use of severity based tempo adaptation followed by autoencoder based speech feature enhancement. All evaluations were carried out on Universal Access dysarthric speech corpus. An overall absolute improvement of 16% was achieved using tempo adaptation followed by autoencoder based speech front end representation for DNN-HMM based dysarthric speech recognition.

Index Terms: Autoencoders, Dysarthric Speech, Tempo adaptation, Speech Enhancement

1. Introduction

Neurological injury or disease such as Amyotrophic lateral sclerosis (ALS), Parkinsons disease (PD) or cerebral palsy resulting in weakness, paralysis, or a lack of co-ordination of the motor-speech system manifests as a speech disorder known as dysarthria. Dysarthria leads to reduction in intelligibility, audibility, naturalness, and efficiency of vocal communication. Owing to the motor impairment, interaction with electronic devices using speech is more effective than through keyboard input [1]. Inter-speaker and intra-speaker inconsistencies in the acoustic space as well as the sparseness of data poses a serious challenge in building automatic speech recognition engine (ASR) system for dysarthric speech. Speaker adaptation based ASR systems and dysarthric speech enhancement to match the characteristics of normal speech are two popular techniques that have been employed to address this challenge.

In [2], a similarity measure between dysarthric speakers to select relevant speaker data for training rather than speaker independent acoustic models, followed by maximum a posteriori (MAP) adaptation has been used. In [3] a more suitable prior model for adaptation based on the dysarthric speaker's acoustic characteristics has been used to achieve improved recognition. ASR accuracy was shown to improve by representing dysarthric speech in terms of articulatory models in [1, 4, 5]. In [6], a set of MFCC features, that best represent dysarthric acoustic features was selected to be used in Artificial Neural Network (ANN)-based ASR. A hybrid adaptation using maximum likelihood linear regression (MLLR) and MAP [7] have been used to improve dysarthric speech recognition. Voice parameters such as jitter and shimmer features along with a multi-taper spectral estimation have been used along with feature space maximum likelihood linear regression (fMLLR) transformation and speaker adaptation to obtain improved dysarthric speech recognition [8].

In [9], the modifications to prosody, spectral content, regions of the signal containing formants, and effects of signal processing on dysarthric speech have been studied. Transformations of dysarthric speech in both the temporal as well as spectral domain have been employed so as to match the characteristics of normal speech. In another study [10], transformations in the temporal domain by adjusting the tempo of speech using phase vocoding, spectral domain transformation using anchor-based morphing of the spectrum and phoneme level correction of pronunciation were used to give improved intelligibility and was validated both by human listeners and ASR based recognition. In [11], vowel space transformations by manipulating vowel duration and formants F1 - F3 stable points were shown to improve the intelligibility of dysarthric speech. In their work [12], authors use speech synthesis to produce utterances with improved intelligibility corresponding to a dysarthric utterance using the dysarthric speaker characteristics. Yet another aspect that has been used to improve ASR performance is based on the severity of the Dysarthria. Traditionally, speech intelligibility has been an indicator of severity of the speech disorder [13]. An understanding of severity has contributed to improved speech recognition of dysarthric speech as seen in [7, 14, 15].

Deep Autoencoder (DAE) based feature enhancement technique provides significant performance gain for speech recognition. A variant of basic DAE, deep denoising autoencoders (DDA) have been used to enhance speech features especially in noisy conditions [16, 17, 18]. DDAs are also efficiently used for reverberant speech recognition [19]. In [20] a DDA is pretrained as restricted Boltzmann machines (RBMs) and then a nonlinear mapping from noisy to clean features is learned from corresponding clean speech features. Generally, a DDA learns a stochastic mapping from noisy to clean by using clean features for fine tuning.

In this paper, we train the deep autoencoder network using healthy control speech which is in turn used to enhance the speech features of dysarthric speech. We propose a method to improve the recognition of dysarthric speech using enhanced speech features that have been extracted using a Deep Autoencoder (DAE). Additionally, we extend our earlier work [21], wherein we transform the dysarthric speech in the temporal domain using severity-based tempo adaptation (TA) and use the tempo adapted dysarthric speech prior to feature enhancement using a DAE. We analyse the contribution of the individual techniques towards improvement in speech recognition as well as tempo adaptation and DAE-based feature enhancement in tandem. To the best of our knowledge, autoencoder-based

http://dx.doi.org/10.21437/Interspeech.2017-1318

speech feature enhancement for dysarthric speech has not been attempted so far and is the main contribution of this paper.

The rest of the paper is organized as follows. Section 2 describes the methodology employed to enhance speech features for dysarthric speech recognition, Section 3 discusses the various experimental setups and a description of the data used, Section 4 describes the results and analysis and we conclude in Section 5.

2. Speech Feature Enhancement

In this paper, we propose (a) an improved front-end speech processing through enhanced speech features using deep autoencoders (DAE) and (b) a combination of dysarthric speech transformation in the temporal domain followed by features enhancement using DAE. Figure 1 shows an overview of the proposed setup for improved dysarthric speech recognition.

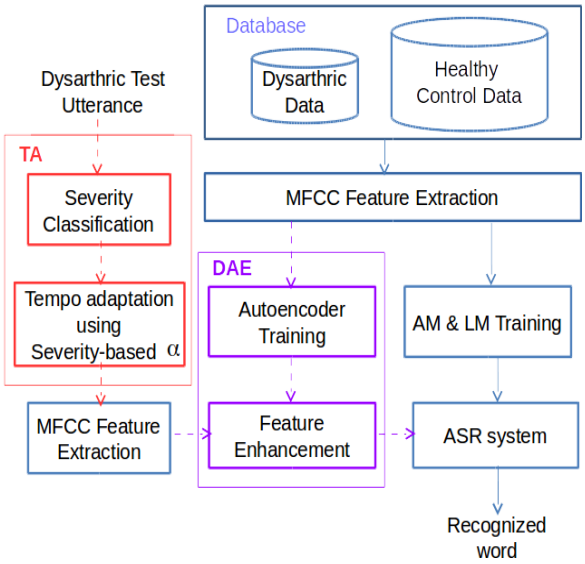

Figure 1: *Proposed setup for improved dysarthric speech recognition*

2.1. Deep Autoencoder (DAE)

Traditionally, an autoencoder is a fully connected artificial neural network system, with a bottleneck layer as shown in Figure 2. In this paper we use deep autoencoder to enhance the Mel Frequency Cepstral Coefficients (MFCC) based features of dysarthric speech.

An autoencoder comprises two blocks, the encoder and the decoder. The objective of the encoder is to transform a higher dimensional input feature vector into a lower dimensional representation at the bottleneck layer. The bottleneck features are then transformed into a higher dimensional representation at the decoder end of the autoencoder, the input and output features drive the learning of the autoencoder, to ensure that the bottleneck layer presents a lower dimensional representation of the input features. Encoding operation can be represented as

$$\boldsymbol{y} = f(\theta; \boldsymbol{x}) = s(W\boldsymbol{x} + \boldsymbol{b}) \quad (1)$$

where

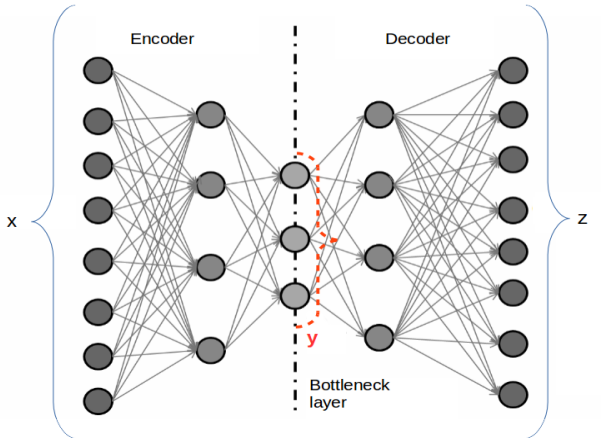

Figure 2: *Deep Autoencoder (DAE)*

- $\boldsymbol{y}$ is the bottleneck feature vector representation of the input feature vector x, which propagates through hidden layers.
- θ = {W, $\boldsymbol{b}$}, where W and $\boldsymbol{b}$ are the weights and biases of the network respectively.
- s is an activation function, linear or non-linear.

At the decoder, the bottleneck feature vector $\boldsymbol{y}$ which propagates through hidden layers is mapped to the higher dimensional representation $\boldsymbol{z}$ at the output stage as

$$z = g(\theta'; \boldsymbol{y}) = s(W'\boldsymbol{y} + \boldsymbol{b'}) \quad \text{where} \quad \theta' = \{W', \boldsymbol{b'}\} \quad (2)$$

Thus, the output of the DAE can be represented as a function of the weights and biases of the encoder and decoder stages, namely $\{\theta\ , \theta'\}$ and written as $\boldsymbol{z} = g(\theta'; (f(\theta; \boldsymbol{x}))$. DAE parameters θ and θ' are optimized such that $\boldsymbol{z}$ is as close as possible to input/target $\boldsymbol{x}$ and maximizes $P(\boldsymbol{x}|\boldsymbol{z})$. The autoencoder parameters are optimized using mean square error (MSE) backpropagation between target $\boldsymbol{x}$ and network output $\boldsymbol{z}$.

2.2. Unsupervised feature extraction using modified DAE

Unsupervised feature learning is currently being used as an alternative to the conventional MFCC features. In this paper, we modify the DAE architecture to suit the purpose of enhancing dysarthric speech features as shown in Figure 3. The DAE parameters ($\theta 1$, $\theta 2$ and θ') are learned from healthy control speech. We have used MFCC features from healthy control speech as input and target, as shown in Figure 3(a). Learned parameters ($\theta 1$, $\theta 2$ and θ') represent the weights and biases of the DAE which provides the minimum MSE between the input and target at the time of autoencoder training. $\theta 1$, $\theta 2$ are encoder parameters and θ' is the decoder parameter of the network where $\theta 1 = \{W1, b1\}$, $\theta 2 = \{W2, b2\}$ and $\theta' = \{W', b'\}$. We extract enhanced features from dysarthric speech using the trained autoencoder parameters. We use these enhanced features as input to the decoding process.

2.3. Severity based Tempo Adaptation (TA)

We examined the improvement in dysarthric speech recognition using severity-based tempo adaptation in one of our earlier works [21]. Dysarthric speech severity level classification was carried out using techniques mentioned in [22]. Malfunctioning of the motor nervous system impacts the precision and

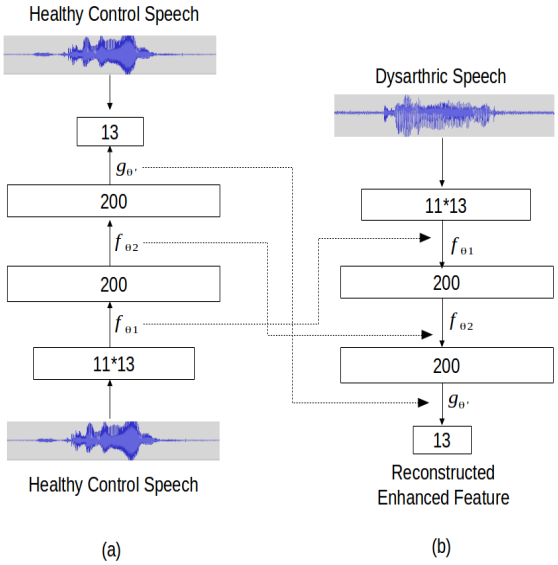

Figure 3: *(a) Modified DAE architecture for Training (learning parameter $\theta 1$, $\theta 2$and θ') (b) Feature extraction for dysarthric speech using learned DAE parameters*

flexibility of the vocal folds, articulators and other speech subsystems, leading to reduced prosodic control. This manifests as longer duration for sonorants in dysarthric speech as compared to normal healthy speech [10]. Temporal reduction of sonorant regions emerges as a possible enhancement to dysarthric speech to provide improved intelligibility, both to human listeners as well as the ASR systems. This process is referred to as tempo adaptation. Tempo adaptation based on the knowledge of severity of the dysarthric speech was found to be beneficial since the adaptation parameter α could be learned for a specific severity level, empirically using healthy control speech data and dysarthric speech of various severity levels, where exact the same words are spoken by both healthy control speakers and dysarthric speakers. Consider a spoken word whose average sonorant duration for healthy control speakers is d_{hc} and that for a dysarthric utterance is d_{dys}. The tempo adaptation parameter for the each word is computed as

$$\alpha = \frac{d_{hc}}{d_{dys}} \tag{3}$$

An average tempo adaptation parameter was computed for each speaker and it was found that tempo adaptation can be carried out by selecting an α value that would suit all the speakers at a certain severity level. Tempo adaptation needs to be carried out in a manner so as to not affect the pitch of the sonorant regions of dysarthric speech. A phase vocoder based on short-time Fourier transform (STFT) is used [23].

Let $X(F)$ be the Fourier transform of a speech signal $x(t)$, $x(t) \overset{\mathscr{F}}{\longleftrightarrow} |X(F)| \cdot \Theta$, where $|X(F)|$ is the magnitude and $\Theta = \angle X(F)$ is the phase. Magnitude spectrum and phase of the STFT are either interpolated or decimated based on the adaptation parameter (α), where the magnitude spectrum is directly used from the input magnitude spectrum and phase values are chosen to ensure continuity. This ensures that the pitch of the time-warped sonorant region is intact. For the frequency band at frequency f and frames i and $j > i$ in the modified spectrogram, the phase Θ is predicted as

$$\Theta'_{jf} = \Theta^f_j + 2\pi f \cdot (i - j) \tag{4}$$

If the modified magnitude and phase spectrum are represented as $|X'(F)|$ and $\angle\Theta'$, the spectrogram is then converted into a time-domain signal using inverse Fourier transform, wherein the tempo of the sonorant regions are adapted with the pitch unchanged as $|X'(F)| \cdot \Theta' \overset{\mathscr{F}^{-1}}{\longleftrightarrow} x'(t)$

Additionally, we explore the possibility of using severity-based tempo adaptation in tandem with DAE-based feature enhancement as shown in Figure 1.

3. Experimental Setup

Data from Universal Access (UA) speech corpus [24] was used for both training and testing. UA dysarthric speech corpus comprises data from 13 healthy control (HC) speakers and 15 dysarthric (DYS) speakers with cerebral palsy. Three blocks of data were collected for each speaker such that in each block a speaker recorded 10 digits, 26 international radio alphabets, 19 computer commands, 100 common words and 100 uncommon words such that each speaker recorded 455 distinct words and a total of 765 isolated words. Speech intelligibility ratings for each dysarthric speaker, as assessed by five naive listeners are also included in the corpus. Based on this evaluation, speakers were divided into four different categories. We have used this information to analyze the performance of our recognition systems at different dysarthria severity level.

Tempo adaptation parameters as shown in Table 1, were empirically determined for different severity levels in the UA speech corpus as described in [21].

Table 1: *Tempo adaptation parameter α based on severity*

Severity	Very Low	Low	Mid	High
α	1.0	0.6	0.5	0.4

We use the Kaldi [25] toolkit-based deep autoencoder for our experiments. The architecture of deep autoencoder (DAE) was 143-200- 200-13, with 143 nodes in the input layer, where 13 dimensional MFCC with a splicing of 11 contextual frames, 200 neurons in each hidden layer and 13 nodes in the output layer. All neurons had sigmoid activation in all the layers. To demonstrate the ability of autoencoder to capture general spectral information, autoencoder was trained using training data as mentioned in Table 2 for each of the four configurations.

3.1. Speech recognition

Kaldi toolkit [25] was used for DNN-HMM based dysarthric speech recognition. The system was trained using a maximum likelihood estimation (MLE) training approach along with 100 senones and 8 Gaussian mixtures. Cepstral mean and variance normalization (CMVN) was applied on each of the above sets of features. Dimensionality reduction was done using Linear Discriminant Analysis (LDA), wherein LDA builds HMM states using feature vectors with a reduced feature space. We use a context of 6 frames (3 left and 3 right) to compute LDA. The feature vector size post LDA is set to 40.

The input layer of DNN has 360 (40×9 frames) dimensions using a left and right context of 4 frames. The output layer has a dimension of 96 (number of senones available in the

data). Two hidden layers with 512 nodes in each layer were used. Dysarthric speech recognition was carried out by using a constrained Language Model (LM), wherein we restrict the recognizer to give one word as output per utterance.Performance of each of the recognition systems is reported in terms of word error rate (WER).

A specific combination of healthy control (HC) and dysarthric data (DYS) from each of the three blocks (B1, B2 and B3) of computer command (CC) words, were used for various experiments as described in Table 2 to prove the feasibility of the proposed method.

Table 2: *Training and testing setup.*

System	Training	Testing
S-1	HC-CC (B1,B3)	HC-CC (B2)
S-2	HC-CC (B1,B3)	DYS-CC (B2)
S-3	DYS-CC(B1,B3)	DYS-CC (B2)
S-4	HC-CC(B1,B3) + DYS-CC(B1,B3)	DYS-CC (B2)

4. Experimental Results

We examine the effectiveness of two types of enhancements to dysarthric speech for automatic speech recognition purpose, namely (1) Tempo adaptation carried out in the temporal domain (2) DAE based MFCC feature enhancement. DNN-HMM based speech recognition was carried out for both the above scenarios individually and in tandem. The DAE and DNN-HMM systems were configured and trained as described in Section 3. ASR performance is reported in terms of word error rates (WERs). The following four different front-end scenarios were considered for our experiments :

- MFCC features
- Tempo adaptation followed by MFCC feature extraction.
- DAE enhanced MFCC features.
- Tempo adaptation followed by DAE enhanced MFCC features.

WERs for each configuration in Table 2 for the relevant front-end scenarios described above can be seen in Table 3. Purpose of S-1 is to examine the impact of DAE on clean or healthy control speech. The WERs for MFCC and MFCC-DAE indicate that DAE-based speech feature enhancement has improved the recognition performance even for healthy-control or clean speech. Significant improvements were seen for all four configurations over the baseline MFCC-based ASR system when enhancements were applied. Although the tandem system showed significant improvement over the baseline (of the order of 16% for S-2) for all configurations, for S-4 the MFCC-DAE seemed to perform the best. When additional dysarthric data was included to the S-2 configuration for training the DAE and DNN-HMM systems, the performance (of S-4) significantly improved across all front-end scenarios. However, the individual front-ends performed on par or slightly better than the tandem front-end. In order to understand this better, we analyze the performances of S-2 and S-4 by looking at the performances of individual and tandem scenarios at dysarthria severity levels as shown in Table 4.

The tempo adaptation parameter used for very low severity was 1, indicating no adaptation is performed on this set of dysarthric speech. Hence we only report the MFCC-DAE performance. The ASR performance across all front-end scenarios

Table 3: *WER for different Experimental setups.*

System	MFCC (Baseline)	TA-MFCC	MFCC-DAE	TA-MFCC + DAE
S-1	2.26	-	0.00	-
S-2	46.89	44.25	34.51	30.71
S-3	32.80	-	27.85	-
S-4	31.59	21.30	20.14	20.69

Table 4: *WER analysis at severity level.*

System	Severity	MFCC (Baseline)	TA-MFCC	MFCC-DAE	TA-MFCC + DAE
S-2	Very-low	14.59	-	2.86	-
	Low	43.79	39.27	14.41	15.54
	Mid	67.63	60.53	60.00	48.16
	High	82.06	80.38	78.71	71.29
S-4	Very-low	12.93	-	1.65	-
	Low	22.60	16.95	13.56	17.23
	Mid	34.47	15.79	14.47	15.79
	High	66.27	61.24	60.29	58.61

reduces with the increase in severity. In majority of the cases, MFCC-DAE provided the best performance or least WER. Addition of dysarthric speech to the training data has given tremendous improvement in the overall performance of S-2 configuration. However, majority of the contribution to this spike in performance comes from the performance improvement for mid and high severity dysarthric speech. Based on the severity level assessment, the tandem system performs best for mid and high severity dysarthric speech while MFCC-DAE gives significant performance gains in case of very low and low severity dysarthric speech. Several iterations with various combinations of data need to be conducted to arrive at an exact recommendation regarding the choice of front-end. However, the tandem system (TA-MFCC+DAE) performed the best or on par with MFCC-DAE in most cases.

5. Conclusions

The objective of this paper was to improve dysarthric speech recognition by enhancing the MFCC-based speech front end. We used deep autoencoders to enhance the Mel Frequency Cepstral Coefficients (MFCC) based features in order to improve dysarthric speech recognition. Additionally, we analyzed the use of severity-based tempo adaptation followed by autoencoder based speech feature enhancement. tempo adaptation was done in the temporal domain using a severity based parameter to match the dysarthric speech to healthy-control speech. Performance of a DNN-HMM speech recognizer for both the enhancement techniques individually as well as in tandem was analyzed. It was observed that each technique provided significant improvement over the baseline recognition. All evaluations were carried out on Universal Access dysarthric speech corpus. An overall absolute improvement of 16% was achieved using tempo adaptation followed by autoencoder based speech front end representation. Further, severity level analysis of the dysarthric recognition provided insights into the choice of front-end for each severity level, wherein the tandem system (TA-MFCC+DAE) performed exceptionally well for mid and high severity levels of dysarthria. Future work could entail optimizations of the DAE network to further improve dysarthric speech recognition.

6. References

[1] F. Rudzicz, "Learning mixed acoustic/articulatory models for disabled speech," in *Proc. NIPS 2010, – Workshop on Machine Learning for Assistive Technologies at the 24th annual conference on Neural Information Processing Systems*, 2010, pp. 70–78.

[2] H. Christensen, I. Casanueva, S. Cunningham, P. Green, and T. Hain, "Automatic selection of speakers for improved acoustic modelling: recognition of disordered speech with sparse data," in *IEEE Spoken Language Technology Workshop (SLT)*, Dec 2014, pp. 254–259.

[3] H. V. Sharma and M. Hasegawa-Johnson, "Acoustic model adaptation using in-domain background models for dysarthric speech recognition," *Computer Speech & Language*, vol. 27, no. 6, pp. 1147 – 1162, 2013, special Issue on Speech and Language Processing for Assistive Technology.

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

[5] S. Hahm, D. Heitzman, and J. Wang, "Recognizing dysarthric speech due to amyotrophic lateral sclerosis with across-speaker articulatory normalization," in *6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2015, p. 47.

[6] S. R. Shahamiri and S. S. B. Salim, "Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach," *Advanced Engineering Informatics*, vol. 28, no. 1, pp. 102 – 110, 2014.

[7] S. Sehgal and S. Cunningham, "Model adaptation and adaptive training for the recognition of dysarthric speech," in *6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2015, p. 65.

[8] C. Bhat, B. Vachhani, and S. K. Kopparapu, "Recognition of dysarthric speech using voice parameters for speaker adaptation and multi-taper spectral estimation," in *In Proc. INTERSPEECH*, 2016, pp. 228–232.

[9] J. P. Hosom, A. B. Kain, T. Mishra, J. P. H. van Santen, M. Fried-Oken, and J. Staehely, "Intelligibility of modifications to dysarthric speech," in *In Proc. ICASSP*, April 2003, pp. I–924–I–927 vol.1.

[10] F. Rudzicz, "Adjusting dysarthric speech signals to be more intelligible," *Computer Speech & Language*, vol. 27, no. 6, pp. 1163 – 1177, 2013, special Issue on Speech and Language Processing for Assistive Technology.

[11] A. B. Kain, J.-P. Hosom, X. Niu, J. P. van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Communication*, vol. 49, no. 9, pp. 743 – 759, 2007.

[12] M. Dhanalakshmi and P. Vijayalakshmi, "Intelligibility modification of dysarthric speech using HMM-based adaptive synthesis system," in *2015 2nd International Conference on Biomedical Engineering (ICoBE)*, March 2015, pp. 1–5.

[13] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "PEAKS A system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, no. 5, pp. 425 – 437, 2009.

[14] M. B. Mustafa, S. S. Salim, N. Mohamed, B. Al-Qatab, and C. Siong, "Severity-based adaptation with limited data for ASR to aid dysarthric speakers," *PLoS One. Jan 23;9(1):e86285. doi: 10.1371/journal.pone.0086285. eCollection*, 2014.

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

[16] P. G. Shivakumar and P. Georgiou, "Perception optimized deep denoising autoencoders for speech enhancement," in *In Proc. INTERSPEECH*, 2016, pp. 3743–3747.

[17] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *In Proc. INTERSPEECH*, 2013, pp. 436–440.

[18] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 1759–1763.

[19] T. Ishii, H. Komiyama, T. Shinozaki, Y. Horiuchi, and S. Kuroiwa, "Reverberant speech recognition based on denoising autoencoder." in *In Proc. INTERSPEECH*, 2013, pp. 3512–3516.

[20] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 1759–1763.

[21] C. Bhat, B. Vachhani, and S. Kopparapu, "Improving recognition of dysarthric speech using severity based tempo adaptation," in *Speech and Computer: 18th International Conference, SPECOM 2016, Budapest, Hungary, August 23-27, 2016, Proceedings*, 2016, pp. 370–377.

[22] ——, "Automatic assessment of dysarthria severity level using audio descriptors," in *In Proc. ICASSP*, March 2017.

[23] M. Portnoff, "Implementation of the digital phase vocoder using the fast fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 243–248, Jun 1976.

[24] 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 *In Proc. INTERSPEECH*, 2008, pp. 1741–1744.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.