SIMULATING DYSARTHRIC SPEECH FOR TRAINING DATA AUGMENTATION IN CLINICAL SPEECH APPLICATIONS

Yishan Jiao[1]*, Ming Tu*[1]*, Visar Berisha*[1,2] *and Julie Liss*[1]

[1]Department of Speech and Hearing Science
[2] School of Electrical, Computer, and Energy Engineering
Arizona State University

ABSTRACT

Training machine learning algorithms for speech applications requires large, labeled training data sets. This is problematic for clinical applications where obtaining such data is prohibitively expensive because of privacy concerns or lack of access. As a result, clinical speech applications typically rely on small data sets with only tens of speakers. In this paper, we propose a method for simulating training data for clinical applications by transforming healthy speech to dysarthric speech using adversarial training. We evaluate the efficacy of our approach using both objective and subjective criteria. We present the transformed samples to five experienced speech-language pathologists (SLPs) and ask them to identify the samples as healthy or dysarthric. The results reveal that the SLPs identify the transformed speech as dysarthric 65% of the time. In a pilot classification experiment, we show that by using the simulated speech samples to balance an existing dataset, the classification accuracy improves by $\sim 10\%$ after data augmentation.

Index Terms— Dysarthric speech, voice conversion, adversarial training, data augmentation

1. INTRODUCTION

Recent studies in machine learning have shown that models built from large data sets can achieve extraordinary performance by mining data-driven features directly from the data. Take large vocabulary continuous speech recognition as an example: databases consisting of thousands of hours of speech from many individuals that cover the large variability in speaking style, environment, speaker age, etc., are required to train powerful deep neural networks (DNNs)-based acoustic models [1].

For consumer applications, speech samples can be collected efficiently on a large scale; however for clinical applications of speech analytics, healthy speech samples have only limited utility. For example, if our aim is to build speech-based assistive technology for patients with amyotrophic lateral sclerosis (ALS), simple application of models trained on healthy speech fail even under moderate dysarthria [2]. Other clinical speech applications that require large labeled training sets include automatic detection of speech disorders [3][4], intelligibility assessment [5][6], automatic recognition of disordered speech [7][8], automated acoustic measures of speech disorders [9][10], etc.

Unlike healthy speech, the collection of pathological speech takes longer to conduct and can be more sensitive to other factors, such as variable recording conditions, uncontrolled body movements, unbalanced samples across speakers and diseases, etc. In the literature, there only exist a few publicly-available datasets that are relatively large (e.g., the Nemours database [11] and the TORGO database [12]); but most researchers opt to collect their own small-scale datasets tailored to their needs. Due to a lack of data, machine learning models used in the study of pathological speech are typically limited to simple unsupervised metrics [13], or flat supervised models [14][15]. When deep learning models are used [16], their solution space is typically constrained using other criteria for better generalization.

Our aim in this paper is to generate simulated dysarthric speech via a model that transforms healthy speech to dysarthric speech so as to augment existing datasets for training large scale machine learning models. To the best of our knowledge, this is the first attempt to augment training data using voice conversion techniques. We restrict our analysis to speech from individuals with ALS, a rapidly progressing neurodegenerative disease. Machine learning models based on speech are particularly useful for this group of patients for building new assistive devices that generalize well across disease conditions and patient speaking styles.

The proposed method includes speaking rate modification using PSOLA, spectral feature transformation using adversarial training, and pitch modification using a linear transformation. We conducted objective and subjective evaluation to examine whether the simulated ALS speech matches true ALS speech in both the acoustic and perceptual domain. Furthermore, we demonstrate the utility of data augmentation on a classification task. In the remaining parts of this paper, Section 2 introduces the proposed transformation framework and experimental settings. The objective and subjective validation of the method are presented in Section 3 along with a pilot experiment of data augmentation using the simulated samples. We conclude with a discussion of our future plans in Section 4.

Relation to previous work: Voice conversion is a technique to modify a source speaker's speech to be perceived as if a target speaker had spoken it [17]. This paper is motivated by this idea, however the source and target speakers in our study are a group of healthy speakers and a group of ALS speakers. Speech transformation of dysarthric speech has also been previously studied [18, 19]. In contrast to these studies, the present study attempts to transform healthy speech to pathological speech for data augmentation rather than doing in the opposite way for improving intelligibility. Adversarial training has been recently explored in the field of voice conversion [20][21][22]. In the most recent work in [20], the authors used variational autoencoding Wasserstein generative adversarial network (VAW-GAN) to transform speech features, with the assumption that there is a latent variable representing the common phonetic content. However, this assumption fails for ALS speech, where there is a significant deviation from healthy speech in the acoustic representation of phonemes. As an alternative, we use deep

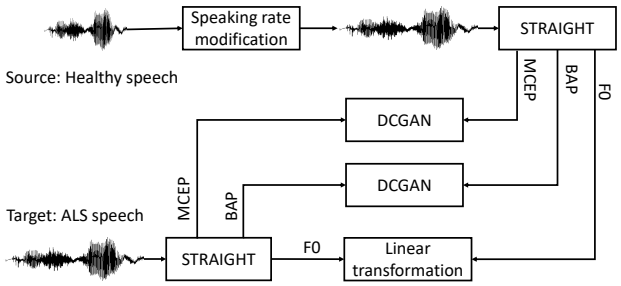

Fig. 1. Proposed transformation frame.

convolutional generative adversarial networks (DCGANs) that has been used for speech synthesis postfiltering [21][22] to transform speech features.

2. PROPOSED METHOD

We begin with the assumption that we have a group of ALS speakers with similar perceptual symptoms (e.g. reduced articulatory precision) we aim to model and a group of healthy speakers with a distinct speaking style (e.g. a regional accent). Our aim is to superimpose the ALS symptoms to the speakers with different speaking styles in an attempt to model the variation induced by the disease *and* the variation in speaking style. This allows us to artificially expand the training set in machine learning applications.

The proposed transformation framework is shown in Figure 1. The source 'speaker' is a group of healthy speakers, and the target 'speaker' is a group of ALS speakers as described above. We assume that speakers from both groups read the same materials and we build a mapping between *each* healthy / ALS speaker pair. Suppose there are N healthy speakers and M ALS speakers. Each speaker reads the same P sentences. The paired samples for training are denoted by $\{S_{i,k}, T_{j,k}\}$, where $S_{i,k}$ and $T_{j,k}$ are the source and target speaker, respectively, with $i = 1, 2, ..., N$, $j = 1, 2, ..., M$, and $k = 1, 2, ..., P$. By performing the training over groups of speakers, we expect that the model will superimpose the ALS symptoms to healthy speech rather than learning a specific speaker's pattern.

People with ALS suffer from mixed flaccid spastic dysarthria, characterized by slow speech rate, imprecise phoneme articulation, hypernasality, monopitch, breathiness, and a harsh voice [23]. To capture these characteristics, we propose the three-step conversion strategy outlined in Figure 1: 1) speaking rate modification; 2) spectral feature transformation using DCGANs; 3) pitch modification. We describe the details of these below.

2.1. Speaking rate modification

The first transformation is to modify the speaking rate of the healthy speech to match the rate of ALS speech. During training, we match each healthy speech sample to the length of its paired ALS speech sample. During testing, each healthy speech sample is modified to a reference value. In our study, the rate of ALS speech was twice as slow on average as the rate of the healthy speech. As a result, for test speech samples, we stretch them to double their lengths. The change in speaking rate is performed by using the open-source Praat Vocal Toolkit [24], which uses PSOLA to change the duration of the speech while preserving the pitch.

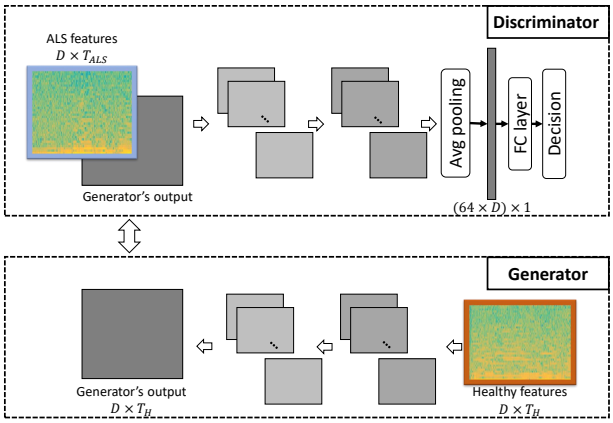

Fig. 2. The structure of DCGAN model used in the presented study.

2.2. Speech feature extraction

As Figure 1 shows, after rate modification, we transform the spectral and pitch features extracted by the STRAIGHT analysis [25]. The fundamental frequency F_0, spectrogram (SP), and the aperiodic spectrum (AP) are extracted from the ALS speech samples and the rate-modified healthy samples. The SP and AP features are then transformed to 39-dimensional mel-cepstral coefficients (MCEPs) and 24-dimensional band-aperiodicity parameters (BAPs), respectively [26]. We expect that the ALS speech characteristics of imprecise phoneme articulation, hypernasality, and breathiness can be modeled by transforming the MCEPs; harsh vocal quality can be modeled by transforming aperiodic components; and monopitch can be modeled by modifying F_0.

2.3. Spectral feature transformation

The spectral features (MCEPs and BAPs) are transformed using adversarial training. Generative adversarial networks (GAN) [27] are a machine learning strategy which uses a combination of discriminative and generative models. The generative model tries to generate samples similar to the target, while the discriminative model tries to distinguish between the distribution of generated samples and actual samples. By training the two models simultaneously, it is expected that the generated samples become indistinguishable from the target sample.

In our study, we take advantage of adversarial training and use DCGANs [28] to transform healthy speech to ALS speech. The structure of the model we used is shown in Figure 2. The lower panel is the generator while the upper panel is the discriminator. The architectures of the multilayer convolutional neural network (CNN) in the generator and the discriminator are shown in Table 1. Within each batch at both the discriminator and the generator, we zero-pad the speech samples until all are of identical temporal length. The input dimension D is the dimension of the spectral features and T_{H} and T_{ALS} are the sequence length (with zero padding) for the healthy and ALS speakers, respectively. Note that T_{H} does not have to equal to T_{ALS} since convolution is used. In both the generator-CNN (G-CNN) and the discriminator-CNN (D-CNN), convolution with padding is used to keep the input and output dimensions the same. For the generator, the input is the healthy speech feature sequence extracted from an utterance (healthy features shown with a red border in the figure), followed by a multilayer CNN (G-CNN). The output of the generator, which has the same size as the generator

Table 1. The structure of DCGAN.

	Generator-CNN	Discriminator-CNN
Input	$D \times T_H$	$D \times T_{ALS}$(or T_H)
Conv1	8 conv with 5×5 kernel size and 1×1 stride	8 conv with 5×5 kernel size and 2×2 stride
Conv2	8 conv with 5×5 kernel size and 1×1 stride	16 conv with 5×5 kernel size and 2×2 stride
Conv3	1 conv with 5×5 kernel size and 1×1 stride	32 conv with 5×5 kernel size and 2×2 stride
Conv4	N/A	64 conv with 5×5 kernel size and 2×2 stride
Output	$D \times T_H$	$64 \times D \times T_{ALS}$(or T_H)

Table 2. Objective and subjective evaluation results

Objective evaluation	***D_p-divergence***	
	Healthy vs. ALS	0.669
	Transformed vs. ALS	0.552
	SVM Classification Results	
	Healthy classified as ALS	2.1%
	Transformed classified as ALS	37.5%
Subjective evaluation	Accuracy on control healthy and ALS samples	98%
	Percentage of transformed perceived as ALS	65%
	Percentage of transformed perceived as ALS with consensus	76%

input, is sent to the discriminator along with the content-parallelled ALS speech feature sequence (ALS features shown with a blue border in the figure). For the discriminator, a multilayer CNN (D-CNN) is connected to the input feature sequence. The D-CNN processes the ALS and transformed speech in the same fashion. Average pooling is used in the D-CNN to process batches with different lengths by temporally averaging the output of the last convolutional layer. Then flattening is applied to concatenate the outputs of average pooling, resulting in a $64D$-dimensional vector. A fully connected layer is used to make the binary classification decision at the discriminator.

Activation functions for the convolution layer are rectified linear unit (ReLU), and the activation function for the fully-connected layer is a sigmoid. The model was trained using Tensorflow [29]. After hyperparameter tuning, the batch size was set to 32, the learning rate was set to 0.00006 with the Adam optimizer [30] with 25 epochs.

2.4. Pitch modification

Pitch is modified using a linear transformation. An important characteristic of ALS speech is reduced pitch variation. We model this reduction in variation through a linear transformation,

$$F_0{}^{\text{trans}}(i) = (F_0(i) - \bar{F}_0) * \alpha + \bar{F}_0 \quad (1)$$

where $F_0(i)$ is the estimated nonzero pitch of the healthy speaker for frame i, and $\alpha = \bar{\sigma}_{F_{0\,ALS}} / \sigma_{F_0}$ is the ratio of standard deviations between the average of ALS speakers and the healthy speaker.

2.5. Experimental settings

We use a subset of a dysarthric speech dataset collected in the Motor Speech Disorders Laboratory at Arizona State University. It consists of speech samples collected from 8 ALS speakers and 8 healthy speakers (4 females and 4 males for each). The dysarthric severity of the ALS speakers ranges from moderate to severe. Each speaker read the same 80 short phrases (6 syllables in each) in English [31]. We used 70 phrases for training and the other 10 for testing. The training data were organized in a speaker independent style within gender, which means that each female/male healthy speaker was mapped to each of the female/male ALS speakers. Therefore, the total number of training samples including both females and males was 2240, and the number of test samples was 80.

During the training stage, we modify the speaking rate of the healthy samples to match their paired ALS speech. Two DCGAN models were trained on the MCEP and BAP features, respectively. The standard deviation of the pitch contour for each ALS speech sample was calculated. During the test stage, the duration of each healthy speech sample was modified to its double length. The MCEPs and BAPs were transformed using the trained DCGAN models, and the F_0 was modified based on Equation 1. STRAIGHT synthesis was used to reconstruct the speech signal.

3. RESULTS

3.1. Objective evaluation

Objective evaluation was performed to examine whether the generated speech is acoustically similar to true ALS speech. First, we compared the distance between the untransformed and the transformed speech features (MCEP, BAP, and F0) of the test speech samples to true ALS speech (ALS speech samples in training data) using the nonparametric D_p divergence measure [32], a measure of distance between two distributions. In order to obtain an unbiased measure, we used bootstrap sampling to ensure the number of samples selected from each of the estimated groups (ALS, healthy, transformed) are the same. The sampling process was done 50 times, and the final measure was obtained by averaging over all trials. The results are shown in Table 2 (D_p-*divergence*). A smaller value of the D_p divergence implies that data from the two classes are more similar. The results show that the transformed speech is more similar to true ALS speech than the healthy speech ($p < 0.01$).

Second, a support vector machine (SVM) was built to distinguish true ALS speech and healthy speech (using samples in the training set) based on the features we developed previously for representing the characteristics of different types of dysarthric speech [33]. The features include: 1) long-term energy spectrum (LTAS), which captures atypical average spectral information in the signal, related to nasality, breathiness, and aypitcal loudness variations of speech; 2) statistics of MFCCs (mean, std, skewness, kurtosis, range, and median absolute deviation); 3) correlation structure features that capture the evolution of vocal tract shape and dynamics at different time scale via auto- and cross- correlation analysis of formant tracks and MFCC. Since speaking rate is an obvious characteristic to separate ALS speech from healthy speech, we excluded features related to rhythm when building the classifier. After training, the classifier was applied to the untransformed test speech samples (healthy) and their transformations. The error rate of the resulting classifier is shown in Table 2 (*SVM Classification Results*). We can see that the model trained on true ALS and healthy speech classifies the transformed samples as ALS 37.5% of the time. In contrast, the untransformed healthy speech samples are only classified as ALS 2.1% of the time.

3.2. Subjective evaluation

Subjective evaluation was performed to examine whether the generated speech was perceptually similar to ALS speech. Five certified SLPs with 12 years clinical experience on average, who routinely work with dysarthric patients, were invited to make a judgement on 20 of the transformed speech samples. The 20 samples were randomly selected from the 80 out-of-sample transformed phrases. The provided instructions were: "Please determine if the speech sample sounds more like ALS or Healthy speech. When you make your choice, please consider if the speaker shows symptoms of ALS."

The generated speech has audible artifacts from the vocoding process. To determine whether the artifacts impact an SLP's decision, we mixed 10 control samples (5 healthy, 5 ALS; vocoded only) with the 20 transformed speech samples. The control samples were generated in the following way: we used the proposed voice conversion framework to transform each control ALS speaker to another ALS speaker, and each control healthy speaker to another healthy speaker. This procedure induces similar artifacts on the control speech samples. The order of the samples was randomized.

For each speech sample, we collected 5 labels from the 5 SLPs and the results were calculated based on all labels on all samples. The results are shown in the subjective evaluation session of Table 2. The first row shows the accuracy of the SLPs perceptual classification on the 10 control healthy and ALS samples (with artifacts induced by the transformation). There was strong consensus on these 10 samples among clinicians, with samples correctly classified 98% of the time. This indicates that the vocoding noises did not impact the ability of the clinicians to correctly classify the speech samples.

The second row shows the percentage of the time that the transformed speech samples were perceived as ALS. However, we noticed that there was consensus (at least 4 same labels) on some of the speech samples, but not on the others. Clinicians also said that it was difficult to make a judgement sometimes. We assumed that for those without consensus, clinicians may make a random guess between ALS and healthy. Therefore, we removed those samples and calculated the result again only when there was consensus (shown in the third row). Across all samples, 65% of the time clinicians classified the transformed speech samples as ALS; for instances where there was consensus among the clinicians, they were classified as ALS 76% of the time. The feedback provided by one of the participating clinicians is as follows:

"*It was a difficult choice for some of them! Some samples sounded disordered, but not necessarily like ALS patients. I defaulted to ALS if rate was slow and I could not get the phonetic content. Some of the features I detected I would describe as slowed rate, altered resonance (hypernasality), breathy vocal quality, and articulatory imprecision. There were some that had an unnatural/robotic/tinny quality that were deviant but not ALS-like.*"

This quote highlights both the benefits and the drawbacks of the proposed method. It is clear that some of the samples exhibit appropriate reductions in articulation precision, unusual resonance, and breathy characteristics; however the transformation method also produces audible artifacts in the speech signal that must be addressed.

3.3. Pilot experiment using data augmentation

To test if the proposed data augmentation method is effective in improving the performance of machine learning models, we designed a pilot experiment to distinguish between ALS speech and ataxic speech. Ataxia is a neurological disorder resulting from damage to the cerebellar control circuit. Ataxic speech has some common characteristics with ALS speech, such as imprecise phoneme articulation, slowed speaking rate, monopitch, and a harsh voice. It also has its own distinct characteristics, including equal and excess stress, irregular articulatory breakdown and excessive loudness variations.

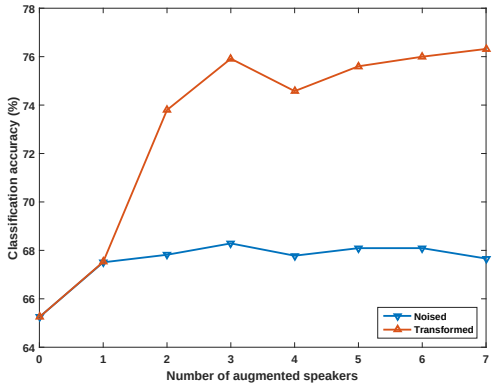

Fig. 3. Classification accuracy after data augmentation.

In our dataset, there are 16 ataxic speakers and 8 ALS speakers with each speaker reading the same 80 phrases. Studies have shown that machine learning classification algorithms are sensitive to unbalanced data. Therefore, we balanced the existing dataset by adding more ALS samples which were transformed from healthy speech using the proposed conversion method. The samples were transformed from the 8 healthy speakers in our dataset.

An SVM model was built based on the same set of features used in the objective evaluation, plus a rhythm feature, the envelope modulation spectrum (EMS), which is a useful indicator of atypical rhythm patterns in pathological speech. The model was trained by iteratively adding one additional simulated speaker to the unbalanced data set. Leave one-speaker (from the original dataset) out cross-validation was used to evaluate the performance. We compared the approach with a more traditional data augmentation method of duplicating the training set speakers by adding noise [34][15]. Here we added white noise (SNR = 10dB) to the training samples. Figure 3 shows the classification accuracies when adding a different number of augmented speakers to the original dataset. We see that the performance gradually improves as additional speakers are added, and the proposed data augmentation method significantly outperformed the augmentation strategy of simply duplicating the training speakers and adding noise. The fluctuations in the accuracy curve in the figure is likely due to the small number of speakers used in the pilot.

4. CONCLUSION

In this paper, we proposed a new data augmentation strategy for clinical speech applications by transforming healthy speech to dysarthric speech. Our objective and subjective evaluation shows that the generated speech are acoustically and perceptually similar to real dysarthric speech, and the pilot classification experiment provides evidence that the augmentation strategy helps improve performance. However, further study is required to determine the benefits of the simulated speech samples in building large scale machine learning models. Future work will focus on reducing the perceptual artifacts by exploring other deep learning models that have been shown to be effective in voice transformation (e.g. recurrent neural networks (RNN)) *and* using the resulting data to build larger machine learning systems.

5. REFERENCES

[1] Dong Yu and Jinyu Li, "Recent progresses in deep learning based acoustic models," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 3, pp. 396–409, 2017.

[2] Ming Tu, Alan Wisler, Visar Berisha, and Julie M Liss, "The relationship between perceptual disturbances in dysarthric speech and automatic speech recognition performance," *The Journal of the Acoustical Society of America*, vol. 140, no. 5, pp. EL416–EL422, 2016.

[3] Everthon Silva Fonseca, Rodrigo Capobianco Guido, Paulo Rogério Scalassara, Carlos Dias Maciel, and José Carlos Pereira, "Wavelet time-frequency analysis and least squares support vector machines for the identification of voice disorders," *Computers in Biology and Medicine*, vol. 37, no. 4, pp. 571–578, 2007.

[4] Björn W Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob Van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, et al., "The interspeech 2012 speaker trait challenge.," in *Interspeech*, 2012, vol. 2012, pp. 254–257.

[5] Jangwon Kim, Naveen Kumar, Andreas Tsiartas, Ming Li, and Shrikanth S Narayanan, "Automatic intelligibility classification of sentence-level pathological speech," *Computer speech & language*, vol. 29, no. 1, pp. 132–144, 2015.

[6] Catherine Middag, Jean-Pierre Martens, Gwen Van Nuffelen, and Marc De Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 3, 2009.

[7] Kristin Rosen and Sasha 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.

[8] Phil D Green, James Carmichael, Athanassios Hatzis, Pam Enderby, Mark S Hawley, and Mark Parker, "Automatic speech recognition with sparse training data for dysarthric speakers.," in *INTERSPEECH*, 2003.

[9] Yishan Jiao, Visar Berisha, Ming Tu, and Julie Liss, "Convex weighting criteria for speaking rate estimation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 23, no. 9, pp. 1421–1430, 2015.

[10] Yishan Jiao, Visar Berisha, and Julie Liss, "Interpretable phonological features for clinical applications," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5045–5049.

[11] Xavier Menendez-Pidal, James B Polikoff, Shirley M Peters, Jennie E Leonzio, and H Timothy Bunnell, "The Nemours database of dysarthric speech," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*. IEEE, 1996, vol. 3, pp. 1962–1965.

[12] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya 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.

[13] Visar Berisha, Julie Liss, Timothy Huston, Alan Wisler, Yishan Jiao, and Jonathan Eig, "Float like a butterfly sting like a bee: Changes in speech preceded Parkinsonism diagnosis for Muhammad Ali," *Proc. Interspeech 2017*, pp. 1809–1813, 2017.

[14] Frank 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.

[15] Ming Tu, Visar Berisha, and Julie Liss, "Objective assessment of pathological speech using distribution regression," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5050–5054.

[16] Ming Tu, Visar Berisha, and Julie Liss, "Interpretable objective assessment of dysarthric speech based on deep neural networks," *Proc. Interspeech 2017*, pp. 1849–1853, 2017.

[17] Alexander Kain and Michael W Macon, "Spectral voice conversion for text-to-speech synthesis," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*. IEEE, 1998, vol. 1, pp. 285–288.

[18] Alexander B Kain, John-Paul Hosom, Xiaochuan Niu, Jan PH van Santen, Melanie Fried-Oken, and Janice Staehely, "Improving the intelligibility of dysarthric speech," *Speech communication*, vol. 49, no. 9, pp. 743–759, 2007.

[19] Frank Rudzicz, "Acoustic transformations to improve the intelligibility of dysarthric speech," in *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*. Association for Computational Linguistics, 2011, pp. 11–21.

[20] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," *arXiv preprint arXiv:1704.00849*, 2017.

[21] Takuhiro Kaneko, Hirokazu Kameoka, Nobukatsu Hojo, Yusuke Ijima, Kaoru Hiramatsu, and Kunio Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *Proc. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2017)*, 2017, pp. 4910–4914.

[22] Takuhiro Kaneko, Shinji Takaki, Hirokazu Kameoka, and Junichi Yamagishi, "Generative adversarial network-based postfilter for STFT spectrograms," in *Proceedings of Interspeech*, 2017.

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

[24] Ramon Corretge, "Praat vocal toolkit: A praat plugin with automated scripts for voice processing," http://www.praatvocaltoolkit.com/, 2012, [Computer software].

[25] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[26] SPTK Working Group et al., "Speech signal processing toolkit (sptk)," *http://sp-tk. sourceforge. net*, 2009.

[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[28] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[29] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[30] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] Julie M Liss, Laurence White, Sven L Mattys, Kaitlin Lansford, Andrew J Lotto, Stephanie M Spitzer, and John N Caviness, "Quantifying speech rhythm abnormalities in the dysarthrias," *Journal of Speech, Language, and Hearing Research*, vol. 52, no. 5, pp. 1334–1352, 2009.

[32] Visar Berisha, Alan Wisler, Alfred O Hero, and Andreas Spanias, "Empirically estimable classification bounds based on a nonparametric divergence measure," *IEEE Transactions on Signal Processing*, vol. 64, no. 3, pp. 580–591, 2016.

[33] Alan Wisler, Visar Berisha, Julie Liss, and Andreas Spanias, "Domain invariant speech features using a new divergence measure," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 77–82.

[34] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al., "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.