INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Few-shot dysarthric speech recognition with text-to-speech data augmentation

Enno Hermann and Mathew Magimai.-Doss

Idiap Research Institute, Martigny, Switzerland
enno.hermann@idiap.ch, mathew@idiap.ch

Abstract

Speakers with dysarthria could particularly benefit from assistive speech technology, but are underserved by current automatic speech recognition (ASR) systems. The differences of dysarthric speech pose challenges, while recording large amounts of training data can be exhausting for patients. In this paper, we synthesise dysarthric speech with a FastSpeech 2-based multi-speaker text-to-speech (TTS) system for ASR data augmentation. We evaluate its few-shot capability by generating dysarthric speech with as few as 5 words from an unseen target speaker and then using it to train speaker-dependent ASR systems. The results indicated that, while the TTS output is not yet of sufficient quality, this could allow easy development of personalised acoustic models for new dysarthric speakers and domains in the future.
Index Terms: automatic speech recognition, dysarthric speech, text-to-speech, few-shot learning

1. Introduction

Dysarthria is a motor speech disorder caused by conditions like Parkinson's disease or amyotrophic lateral sclerosis (ALS). These patients could especially benefit from assistive voice technology, but current ASR systems perform poorly on dysarthric speech due to the differences to typical speech and a scarcity of training data.

Recording large amounts of data can be exhausting for speakers with dysarthria. Few-shot learning approaches, where an acoustic model can be trained with only very little data from a target speaker, are therefore of particular interest.

Few-shot and even zero-shot approaches to pathological speech recognition can be successful [1, 2, 3]. Out of the box, a very large acoustic model with up to 10 billion parameters trained on 4.5 million hours of speech [1] reaches state-of-the-art performance on AphasiaBank [4], a database of aphasic speech. Fine-tuning on this data gives a further 50% relative improvement. However, such amounts of training data are only available to a few private companies. Even fine-tuning and applying a pretrained model with so many parameters is challenging and storing personalised models for each speaker is costly [5]. It is therefore desirable to also investigate more moderately sized models and alternative few-shot approaches.

Voice conversion (VC) is increasingly used as data augmentation for dysarthric speech recognition [6]. A mapping from unimpaired control to dysarthric speakers or between different dysarthric speakers is learned, so that additional speech for ASR training can be generated. This requires that recordings of the target utterances are available. Existing applications to dysarthric ASR have also largely been restricted to VC models that convert only between single pairs of speakers, although in general many-to-many VC approaches also exist [7].

Data augmentation with TTS is an alternative to VC. It allows to synthesise speech for arbitrary sentences and therefore to quickly adapt an ASR system to new commands and domains and a single model can handle any number of speakers. TTS-based data augmentation has already been applied to ASR for low-resource languages and children's speech [8]. ASR and TTS are also naturally linked, corresponding to speech perception and speech production, and joint training in a speech chain has been proposed [9].

In this paper we build upon previous work on TTS for dysarthric speech [10]. They introduced a dysarthria embedding for the FastSpeech 2 TTS system [11] that allows to explicitly model and generate speech of different severity levels. We confirm their finding that data augmentation with synthetic speech is beneficial for dysarthric ASR on a different corpus. We then ask whether dysarthric TTS could also be used to generate ASR training data for a new speaker based on just a small number of recordings. While we find that the synthetic speech on its own is not of sufficient quality to train an ASR system – regardless of whether the speaker has been seen before or not – together with typical speech it works better than typical speech by itself.

2. Methods

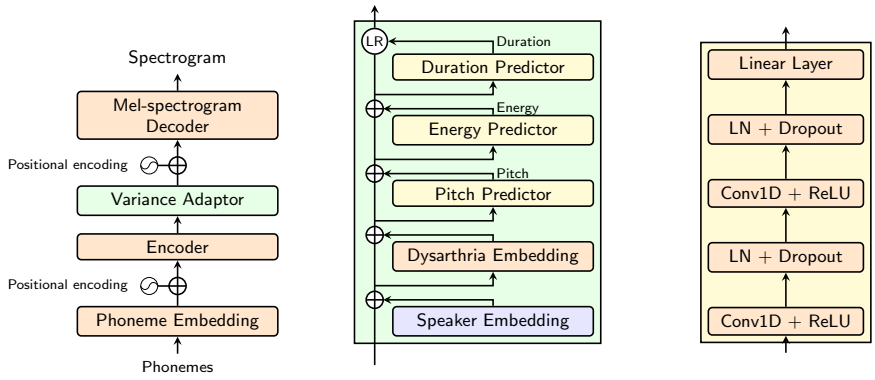
In this section we describe the works on which our dysarthric TTS pipeline is based and any modifications we have made.

2.1. Controllable TTS

FastSpeech 2 [11] is a transformer-based non-autoregressive TTS system that allows for fast training and inference. Figure 1 illustrates the model architecture. It consists of a phoneme encoder and a Mel-spectrogram decoder. In between, it has a *variance adaptor* block to model different sources of variance in the speech signal and to control the TTS output. The variance adaptor contains multiple variance predictors. These are small neural networks that are trained to predict attributes like pitch, energy and phoneme duration. A length regulator expands the encoded input from phoneme- to frame-level based on the durations, while embeddings from the other predictors are added to the input. At training time, ground-truth values are used instead of the predictions.

The original FastSpeech 2 [11] predicts pitch spectrograms obtained from the continuous wavelet transform, but we use an implementation that directly predicts pitch values [12]. We also follow their approach of placing the length regulator after all other variance predictors.

10.21437/Interspeech.2023-2481



(a) *FastSpeech 2* (b) *Variance adaptor* (c) *Variance predictors*

Figure 1: *Our FastSpeech 2 architecture (figure adapted from [11]). LR in subfigure (b) denotes the FastSpeech 2 length regulator and LN in subfigure (c) denotes layer normalisation. Speaker embeddings are obtained from a pretrained model and remain fixed.*

2.2. Multi-speaker TTS

FastSpeech 2 has been extended to multiple speakers by adding a speaker embedding to the encoded input [12]. The following variance predictors are thus conditioned on the speaker identity. The authors found that speaker embeddings from a generative VC system performed better than jointly trained ones or embeddings trained on a discriminative speaker classification task like x-vectors [13]. They chose embeddings from the AdaIN-VC system for one-shot voice conversion [14], so that the TTS would also support speakers not seen during training.

AdaIN-VC [14] is able to convert an utterance to an unseen speaker's voice from a single sample by separately encoding speaker and content. Speaker labels are not required for training, the speaker identity is assumed to be in the constant information throughout an utterance, while the content information is changing. An adaptive instance normalisation (AdaIN) [15] layer means that no parameters have to be learned for a new speaker.

2.3. Dysarthric TTS

Soleymanpour et al. [10] added a dysarthria severity predictor before the other variance predictors, so that their embeddings are conditioned on the severity of dysarthria of the speaker. Due to the controllable nature of FastSpeech 2, speech of different severity levels can then be generated, which they used for data augmentation in a dysarthric ASR system. As severity depends only on the speaker and cannot be predicted from text, we just use a severity embedding and train it with the rest of the model instead of a separate predictor network. We group the speakers into the same 3 groups with their own embedding: unimpaired control speech, mild to moderate dysarthria, severe dysarthria.

They trained speaker embeddings jointly with the FastSpeech 2 model, limiting the set of speakers for which speech can be synthesised to those present in the training data. In this work, we have no such restriction because of the one-shot capable AdaIN-VC speaker embeddings and we investigate how little data is required from a target speaker to synthesise dysarthric speech and build a speaker-dependent ASR system for them. We do not follow their approach of adding heuristics to insert pauses into the synthetic dysarthric speech as we only generate isolated words in this work.

3. Experimental setup

3.1. Datasets

We conducted our study on the UA-Speech [16] database of dysarthric speech. It contains only isolated words, split into 3 blocks, recorded with a 7-microphone array from 15 dysarthric and 13 control speakers without any speech impairment. We use the segmentation of Xiong et al. [17] that removes some excessive silence portions based on forced alignment with a Gaussian mixture model (GMM) ASR system. The dysarthric speech from block 2 of UA-Speech is our test set, which is the standard protocol.

The audio files have a sampling rate of 16 kHz. For compatibility with existing code and pretrained models, we upsample the data to 22050 Hz in the TTS pipeline, while all ASR models are trained on 16 kHz.

3.2. TTS

We use synthetic speech for data augmentation, where we assume that training data for a target speaker is available, and in a few-shot setting, where we apply a trained TTS model on unseen speakers.

For data augmentation, we train one TTS model on all the training data from UA-Speech. For the few-shot experiments, we train 15 different models in a leave-one-speaker-out setup, i.e. on all control and the 14 other dysarthric speakers. We then use different amounts of dysarthric speech from blocks 1 and 3 of UA-Speech to obtain the speaker embeddings and as additional sources of ASR training data.

In each case, we train a phoneme-based FastSpeech 2 TTS model[1] with a batch size of 16 for 500k iterations in the default configuration. The input features are 80-dimensional Mel spectrograms. We obtain ground-truth phoneme durations for the duration predictor from forced alignment with a Kaldi [18] GMM ASR system trained on the same data. Speaker embeddings are from the AdaIN-VC model described in the next section.

For vocoding, we use the pretrained universal HiFi-GAN [19] model[2]. We experimented with fine-tuning the vocoder on UA-Speech, but did not observe consistent benefits. We downsample its 22050 Hz output to 16 kHz for ASR training.

[1] https://github.com/ming024/FastSpeech2

[2] https://github.com/jik876/hifi-gan

3.3. Speaker embeddings

We train AdaIN-VC models[3] on the same data as the TTS models with a batch size of 128 for 200k iterations using the default configuration, also with a leave-one-speaker-out setup. We train on the same Mel spectrograms as for FastSpeech 2 training as in [14]. We take the 128-dimensional output of the speaker encoder as embeddings for FastSpeech 2 training and inference. We do not fine-tune these embeddings during TTS training.

For the few-shot experiments, we select subsets of 5 and 100 words from the UA-Speech training blocks 1 and 3. We do not sample randomly, but instead choose words that offer the broadest phoneme coverage, emulating a scenario where target speakers are asked to record a small list of words with the biggest performance benefit. For each speaker, we pick a random utterance of each word, extract the AdaIN-VC embedding for it and take their average as the speaker embedding for speech synthesis, following Chou et al. [14]. The TTS model is not trained or fine-tuned on these few-shot utterances, although fine-tuning could be explored in the future.

3.4. ASR

All our ASR models are trained with Kaldi [18]. The UA-Speech recipe is adapted from Xiong et al. [17][4]. We train speaker-dependent acoustic models on only the data of the target dysarthric speaker, possibly augmented with synthetic speech. First, a GMM is trained, which serves as a basis for sequence-discriminative lattice-free maximum mutual information (LF-MMI) [20] training of a factorised time-delay neural network (TDNN) [21] acoustic model with 40-dimensional Mel-frequency cepstral coefficients (MFCCs) as input features. Although it is commonly done in LF-MMI training, we do not apply speed perturbation [22] in Kaldi because we can already manipulate the speed during TTS data augmentation.

We decode with a unigram grammar containing only the words from block 2 of UA-Speech as in previous works [17, 23]. In line with those, we group the speakers by severity based on subjective intelligibility ratings included with the corpus as shown in Table 1 and report the word error rate (WER) of each group and the overall WER.

Table 1: *Severity of UA-Speech dysarthric speakers, based on subjective intelligibility ratings (in parentheses).*

Severity	**Speakers**
Severe (0–25%)	M04, F03, M12, M01
Moderate-severe (26–50%)	M07, F02, M16
Moderate (51–75%)	M05, M11, F04
Mild (76–100%)	M09, M14, M10, M08, F05

4. Results

We do not directly evaluate the quality of the synthetic dysarthric speech as we are only interested in its contributions to ASR performance. In the future, it would be worthwhile to apply the objective evaluation measures proposed by Halpern et al. [24]. However, we find that the dysarthria embedding learns to correctly influence the length regulator, with average utterance durations of 1.2s for control, 1.9s for mildly dysarthric and 2.6s for severely dysarthric synthesised speech.

[3]https://github.com/cyhuang-tw/AdaIN-VC

[4]https://github.com/ffxiong/uaspeech

Table 2: *Word error rates (WER) for each group of dysarthric speakers. For clarity, we also indicate whether the target speaker was seen during TTS training or not, where applicable.*

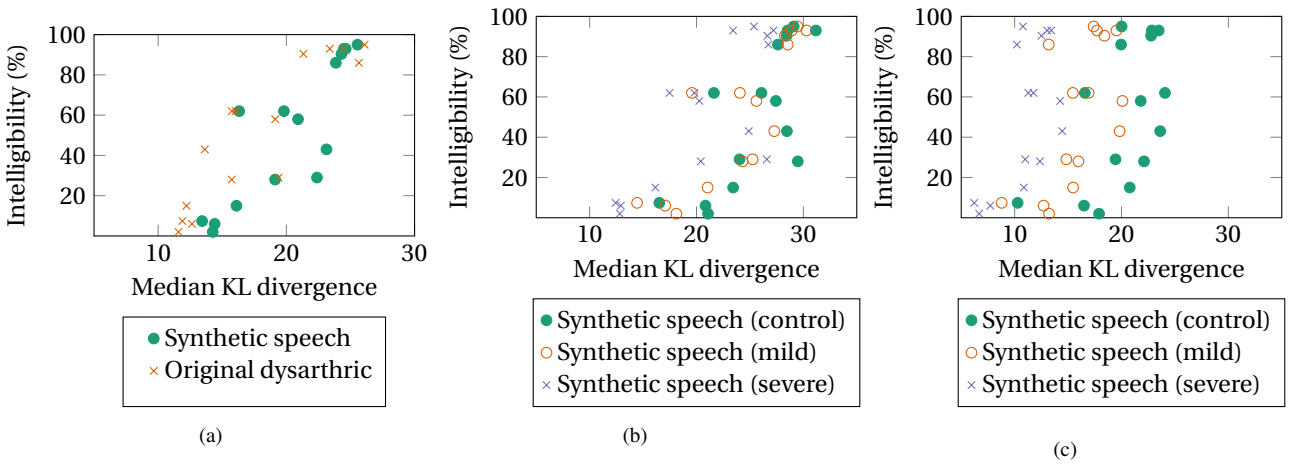
Systems	**Seen**	**Sev.**	**Mod.-sev.**	**Mod.**	**Mild**	**Total**
Baselines						
CTL	-	96.2	74.5	55.1	23.2	56.9
Top-line	-	70.3	42.7	38.2	24.0	41.3
+ CTL	-	65.8	34.3	25.3	15.4	32.8
Data augmentation						
TTS-aug	✓	70.8	38.7	33.6	18.5	37.6
TTS-aug4	✓	68.5	36.9	32.4	19.2	36.7
Few-shot						
F5-ctl	✗	99.6	99.1	98.1	92.0	96.5
+ CTL	✗	94.9	75.9	55.8	22.3	56.7
F100-ctl	✗	98.8	99.0	92.5	83.3	91.7
+ CTL	✗	93.8	75.6	51.7	21.8	55.4
F5-dys	✗	99.4	99.6	98.5	95.4	97.8
+ CTL	✗	94.4	76.1	53.9	22.5	56.8
F100-dys	✗	99.3	99.2	95.6	91.3	95.6
+ CTL	✗	94.5	72.7	52.4	20.6	54.6
F5-mix	✗	99.3	99.1	98.3	92.1	96.5
+ CTL	✗	94.7	75.8	55.6	21.4	56.3
F100-mix	✗	98.6	97.1	92.7	82.6	91.3
+ CTL	✗	93.7	72.7	50.7	20.9	54.2
TTS-only	✓	98.2	93.9	87.8	85.1	90.5
TTS-only4	✓	98.0	92.6	86.5	79.7	87.9

For reference, we show the performance of an ASR system trained only on the control speech (*CTL*) from UA-Speech, see the first row in Table 2. We then train top-line speaker-dependent (SD) systems with all the available dysarthric speech from UA-Speech training blocks 1 and 3. This represents the theoretical upper limit we can reach through data augmentation from a subset of that data. For comparison, we also train SD models that additionally include all control speech (*+CTL*). We note that because we do not use speed perturbation, this top-line does not match the speaker-dependent results of the otherwise similar recipe from Xiong et al. [25].

First, we confirm the findings of Soleymanpour et al. [10] that augmenting the training data with synthetic dysarthric speech (*TTS-aug*) improves speech recognition. We also confirm that adding four times as much synthetic speech further lowers the WER (*TTS-aug4*).

We compare estimating the speaker embedding from 5 (*F5*) and 100 (*F100*) single-word utterances of the target speaker. These utterances are then also included for the training of the acoustic model. In either case, the total number of ASR training utterances is matched with the baseline. All of these models perform poorly with average WERs in the nineties, not even coming close to the control speech model. Nevertheless, we can observe certain patterns, e.g. estimating the speaker embedding from more utterances improves results.

We either set the dysarthria embedding to generate control speech (*F5/100-ctl*), speech of the same severity as the target speaker (*F5/100-dys*) or a mix of control, mild, and severely dysarthric speech (*F5/100-mix*). Curiously, we find that this mix or generating only control speech works better than matching the target severity. This could be because synthesising dysarthric speech introduces some dysarthria-like characteristics that are nonetheless not representative of the target speaker and more detrimental for ASR because the speaker embedding is only



(a) (b) (c)

Figure 2: *Relationship between the median KL divergence of acoustic units and subjective intelligibility ratings of (a) synthetic dysarthric speech used for data augmentation (Pearson's r = 0.85) and of the original dysarthric speech (r = 0.90). (b) shows the same relationship for speech synthesised with the dysarthria embedding set to control, mild, or severe. (c) shows the same relationship for synthetic dysarthric speech used for the few-shot experiments with the dysarthria embedding set to control, mild, or severe. The speaker embeddings are estimated from 5 utterances from the target speaker, but their speech is not included during TTS training.*

designed to capture general speaker information.

We also see slight improvements when combining the F100 data with control speech (+*CTL*). This indicates that while the synthetic speech on its own is not yet of sufficient quality, it can still yield benefits in combination with other data. To further evaluate this, we train another set of SD models on only the synthesised portion of the data used in the TTS-aug experiments, where the target speakers were already seen during TTS training (*TTS-only*). Indeed, even these results are very poor although the speakers were seen and the TTS output was beneficial for ASR data augmentation. This suggests that no significant improvements can be expected in the few-shot setting before the TTS quality in general is not further increased.

5. Analysis

We evaluate the quality of the synthetic dysarthric speech by analysing its acoustic discriminability as proposed in [23]. This approach measures acoustic discriminability by computing Kullback-Leibler (KL) divergences between Gaussian distributions estimated for each acoustic unit (clustered context-dependent triphones) of the ASR system.

Figure 2a shows the relationship between median KL divergences of the synthetic speech used in the data augmentation experiments for each dysarthric speaker and their subjective intelligibility ratings (Pearson's r = 0.85), compared with the original dysarthric speech (r = 0.90). In terms of acoustic space discriminability, the synthetic speech is correctly showing the same patterns as the original dysarthric speech.

For data augmentation, we synthesised speech with the dysarthria embedding set to a different random value for each utterance. But how does the TTS output change when we set the dysarthria embedding to generate control, mild, or severely dysarthric speech? For each embedding value, we synthesise one utterance for each word in the UA-Speech training data. We find that the dysarthria embedding learns to correctly influence the length regulator, with average utterance durations of 1.2s for control, 1.9s for mildly dysarthric and 2.6s for severely dysarthric synthesised speech. Figure 2b shows the relationship between median KL divergences of these three sets of synthesised speech and the subjective intelligibility ratings of each dysarthric speaker. Indeed, the median KL divergences decrease for mild and severely dysarthric synthesised speech, indicating reduced discriminability. We note that when synthesising with the dysarthria embedding set to *control*, there is still a correlation between median KL divergences and subjective intelligibility ratings. This is due to the speaker embedding that inevitably also captures dysarthria characteristics of the speaker, so it is not expected that this synthesised control speech sounds like a control speaker without dysarthria.

However, in the few-shot experiments we synthesised speech for new speakers that were not seen during TTS training. We again generate a set of control, mild, and severely dysarthric speech by setting the dysarthria embedding accordingly with the few-shot model for each unseen speaker. Figure 2c shows that there are meaningful differences in the acoustic space between the three severity levels for these unseen speakers as well.

6. Conclusion

In this paper we confirmed that TTS can be successfully used for data augmentation in dysarthric ASR. However, we found that this method cannot be applied to unseen speakers because the synthetic speech on its own is not of sufficient quality. Possibly, the low number of dysarthric speakers in the training data is not enough to model the significant variability of dysarthric speech. However, we found that the TTS learns to model dysarthric speech characteristics and reproduces differences in acoustic space discriminability between speakers of different severity that are observed in the original dysarthric speech.

In the future, we would like to include a larger set of dysarthric speakers in TTS training to better model their diversity. Similarly, another promising direction would be to train an end-to-end ASR model on multiple dysarthric speakers and then only fine-tune it on the augmented data for a target speaker.

7. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No 766287 (TAPAS) and Innosuisse grant agreement No PFFS-21-47 (IICT flagship).

8. References

[1] A. Xiao, W. Zheng, G. Keren, D. Le, F. Zhang, C. Fuegen, O. Kalinli, Y. Saraf, and A. Mohamed, "Scaling ASR Improves Zero and Few Shot Learning," Tech. Rep. arXiv:2111.05948, 2021.

[2] J. R. Green, R. L. MacDonald, P.-P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, M. A. Ladewig, J. Tobin, M. P. Brenner, P. C. Nelson, and K. Tomanek, "Automatic Speech Recognition of Disordered Speech: Personalized Models Outperforming Human Listeners on Short Phrases," in *Proc. Interspeech*, 2021, pp. 4778–4782.

[3] J. Tobin and K. Tomanek, "Personalized Automatic Speech Recognition Trained on Small Disordered Speech Datasets," in *Proc. ICASSP*, 2022, pp. 6637–6641.

[4] B. MacWhinney, D. Fromm, M. Forbes, and A. Holland, "AphasiaBank: Methods for studying discourse," *Aphasiology*, vol. 25, no. 11, pp. 1286–1307, 2011.

[5] K. Tomanek, V. Zayats, D. Padfield, K. Vaillancourt, and F. Biadsy, "Residual Adapters for Parameter-Efficient ASR Adaptation to Atypical and Accented Speech," in *Proc. EMNLP*, 2021, pp. 6751–6760.

[6] L. Prananta, B. M. Halpern, S. Feng, and O. Scharenborg, "The Effectiveness of Time Stretching for Enhancing Dysarthric Speech for Improved Dysarthric Speech Recognition," in *Proc. Interspeech*, 2022, pp. 36–40.

[7] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "StarGAN-VC2: Rethinking Conditional Methods for StarGAN-Based Voice Conversion," in *Proc. Interspeech*, 2019, pp. 679–683.

[8] V. Kadyan, H. Kathania, P. Govil, and M. Kurimo, "Synthesis Speech Based Data Augmentation for Low Resource Children ASR," in *Speech and Computer*, A. Karpov and R. Potapova, Eds. Springer, Cham, 2021, vol. 12997, pp. 317–326.

[9] A. Tjandra, S. Sakti, and S. Nakamura, "Listening while speaking: Speech chain by deep learning," in *Proc. ASRU*, 2017, pp. 301–308.

[10] M. Soleymanpour, M. T. Johnson, R. Soleymanpour, and J. Berry, "Synthesizing Dysarthric Speech Using Multi-Speaker Tts For Dysarthric Speech Recognition," in *Proc. ICASSP*, 2022, pp. 7382–7386.

[11] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in *Proc. ICLR*, 2021.

[12] C.-M. Chien, J.-H. Lin, C.-y. Huang, P.-c. Hsu, and H.-y. Lee, "Investigating on Incorporating Pretrained and Learnable Speaker Representations for Multi-Speaker Multi-Style Text-to-Speech," in *Proc. ICASSP*, 2021, pp. 8588–8592.

[13] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, 2018.

[14] J.-C. Chou and H.-Y. Lee, "One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization," in *Proc. Interspeech*, 2019, pp. 664–668.

[15] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. ICCV*, 2017, pp. 1501–1510.

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

[17] F. Xiong, J. Barker, and H. Christensen, "Phonetic Analysis of Dysarthric Speech Tempo and Applications to Robust Personalised Dysarthric Speech Recognition," in *Proc. ICASSP*, 2019, pp. 5836–5840.

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

[19] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Proc. NeurIPS*, 2020.

[20] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI," in *Proc. Interspeech*, 2016, pp. 2751–2755.

[21] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in *Proc. Interspeech*, 2018, pp. 3743–3747.

[22] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio Augmentation for Speech Recognition," in *Proc. Interspeech*, 2015, pp. 3586–3589.

[23] E. Hermann and M. Magimai-Doss, "Handling acoustic variation in dysarthric speech recognition systems through model combination," in *Proc. Interspeech*, 2021, pp. 4788–4792.

[24] B. M. Halpern, J. Fritsch, E. Hermann, R. van Son, O. Scharenborg, and M. M. -Doss, "An Objective Evaluation Framework for Pathological Speech Synthesis," in *Proc. ITG Conference on Speech Communication*, 2021.

[25] F. Xiong, J. Barker, Z. Yue, and H. Christensen, "Source Domain Data Selection for Improved Transfer Learning Targeting Dysarthric Speech Recognition," in *Proc. ICASSP*, 2020, pp. 7424–7428.