



This is a repository copy of *Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/213565/>

Version: Published Version

Proceedings Paper:

Leung, W.-Z. orcid.org/0009-0003-4888-1951, Cross, M., Ragni, A. et al. (1 more author) (2024) Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis. In: Proceedings of Interspeech 2024. Interspeech 2024, 01-05 Sep 2024, Kos island, Greece. International Speech Communication Association (ISCA) , pp. 2494-2498.

<https://doi.org/10.21437/Interspeech.2024-1645>

© 2024 ISCA. Reproduced in accordance with the publisher's self-archiving policy.

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>



Training Data Augmentation for Dysarthric Automatic Speech Recognition by Text-to-Dysarthric-Speech Synthesis

Wing-Zin Leung¹, Mattias Cross¹, Anton Ragni¹, Stefan Goetze¹

¹Speech and Hearing (SPandH), Dept. of Computer Science, The University of Sheffield, UK

{WLeung5, MCross2, a.ragni, s.goetze}@sheffield.ac.uk

Abstract

Automatic speech recognition (ASR) research has achieved impressive performance in recent years and has significant potential for enabling access for people with dysarthria (PwD) in augmentative and alternative communication (AAC) and home environment systems. However, progress in dysarthric ASR (DASR) has been limited by high variability in dysarthric speech and limited public availability of dysarthric training data. This paper demonstrates that data augmentation using text-to-dysarthric-speech (TTDS) synthesis for finetuning large ASR models is effective for DASR. Specifically, diffusion-based text-to-speech (TTS) models can produce speech samples similar to dysarthric speech that can be used as additional training data for fine-tuning ASR foundation models, in this case Whisper. Results show improved synthesis metrics and ASR performance for the proposed multi-speaker diffusion-based TTDS data augmentation for ASR fine-tuning compared to current DASR baselines.

Index Terms: Dysarthric speech recognition, diffusion, text-to-speech synthesis, data augmentation

1. Introduction

Dysarthria is a type of motor speech disorder (MSD) that reflects abnormalities in motor movements required for speech production [1]. The psychosocial impact [2] and restrictions on functioning and participation [3] for PwD are well documented [4, 5]. DASR has important implications for AAC devices and home environmental control systems [6, 7]. Although the accuracy of ASR systems for typical speech has improved significantly [8, 9], there are challenges inherent with DASR. Due to high inter- and intra-speaker variability in dysarthric speech and limited public availability of dysarthric data, generic ASR models usually do not generalise well to dysarthric speakers [10, 11]. Baseline ASR models [12], typically trained on larger amounts of typical speech data, can be adapted to domains with limited data availability, such as dysarthric speech recordings [13]. Model adaptation approaches show improved performance [14, 15] and deep learning in combination with data augmentation techniques to address data sparsity [16] have achieved state-of-the-art (SotA) performance for DASR.

Transformer-based models have not been adequately explored for dysarthric speech as such architectures require a significant amount of training data that is not publicly available. [16] implements a spatial convolutional neural network (CNN) with multi-head Transformers (pre-trained on control

speaker data) to recognise visual representations of whole-word dysarthric speech, and [17] use dysarthric and typical speech corpora with data augmentation techniques to implement two-step parameter adjustments to train a dysarthric Transformer model. A recurrent neural network (RNN)-Transducer model has been trained on the Euphonia dataset [18], which contains over 1400 hours of audio data, however, the dataset is not publicly available. Representations from foundation ASR models have been used as input features for DASR systems (e.g. Wav2Vec2.0 [19] and WavLM [20]), but foundation models have not yet been adapted for DASR.

Data augmentation techniques have been widely studied for typical speech tasks [21], but data augmentation for DASR requires further research [22]. Spectro-temporal differences between typical and dysarthric speech (e.g. speaking rate) have influenced approaches, such as vocal tract length perturbation (VTLP) [23], tempo-stretching [24] and speed perturbation [25]. Although slower speaking rates and modifications to the spectral envelope can be modelled, perceptual dysarthric speech characteristics (e.g. articulatory imprecision or voice quality [26]) are not captured. Subsequently, generative adversarial networks (GANs) have been applied to speed-perturbed typical speech for speech synthesis [22] and voice conversion (VC) [27]. Also, Transformer-based systems have been implemented for TTDS synthesis [28, 29]. Recently, diffusion probabilistic modelling (DPM) has been applied to a VC task for dysarthric data augmentation [30]. The results demonstrate improved word error rate (WER) performance of DASR systems using augmented data, and subjective evaluations by human expert listeners show that severity characteristics of dysarthric speech are captured in the synthesis. This paper proposes (i) to create DASR training data by DPM, training Grad-TTS [31] from scratch on dysarthric data to (ii) analyse the use of additional augmented data only to finetune large ASR models (here Whisper), i.e. without matched control speaker data. The remainder of the paper is structured as follows: Section 2 describes the TTDS system and the ASR model adaptation. Experiments are described in Section 3 and their results are presented in Section 4. Section 5 concludes the paper.

2. Methodology

2.1. Dysarthric Speech Synthesis

To synthesise dysarthric mel-spectrogram data \mathbf{X} for DASR augmentation, we train the Grad-TTS [31] model.¹

$$\mathbf{X} = \text{Grad-TTS}(\mathbf{y}, s_E) \quad (1)$$

¹Grad-TTS code adapted from <https://github.com/huawei-noah/Speech-Backbones>. Implementation available at <https://github.com/WingZLeung/TTDS>.

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

from scratch on dysarthric data to augment the limited available dysarthric speech data for a speaker identity s_E . Grad-TTS [31] produces mel-spectrograms matrices \mathbf{X} of size $F \times L$, with $F = 80$ in this work, and L being the variable number of frames. This starts by forming initial distributions centred on the Grad-TTS text-encoder output $\mathbf{E}_\theta(\mathbf{y})$ [31] with trainable parameters θ for a given text-sequence vector \mathbf{y} , and applying a diffusion process (cf. (4)) to denoise the initial distribution of the noisy mel-spectrogram matrix $\mathbf{X}_T \sim \mathcal{N}(\mathbf{E}_\theta(\mathbf{y}), \mathbf{I})$ into an estimate of a training-distribution sample matrix \mathbf{X}_0 , with \mathbf{I} being the identity matrix, and $T = 1.0$. The text encoder \mathbf{E}_θ has the same transformer architecture and training objective as Glow-TTS [32]. The sampling process from \mathbf{X}_T to \mathbf{X}_0 (cf. 4) is defined as the reverse of a forward diffusion process from \mathbf{X}_0 to \mathbf{X}_T with change in \mathbf{X} denoted as

$$d\mathbf{X} = \frac{1}{2}(\mathbf{E}_\theta(\mathbf{y}) - \mathbf{X})\beta(t)dt + \sqrt{\beta(t)}d\mathbf{W}_t \quad (2)$$

where \mathbf{W} is a random process. The forward process in (2) is a continuous mean-reverting variance-preserving stochastic differential equation (SDE). The linear noise schedule

$$\beta(t) = \beta_0 + (\beta_T - \beta_0)t \quad (3)$$

starts at β_0 and ends at β_T to control the perturbation gradient [31]. Intuitively, $\beta(t)$ increases the amount of noise perturbation of \mathbf{X} as t increases. The amount of noise added to \mathbf{X} at each time step dt increases proportionally with t . The encoder term $\mathbf{E}_\theta(\mathbf{y})$ ensures that the distribution of \mathbf{X} is centered on the text-encoder prediction for any t . Default values for Grad-TTS are $\beta_0 = 0.05$ and $\beta_T = 20$ [31], the choice is inspired by diffusion image synthesis [33] where $\mathbf{X}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This work explores reduced β_T to reflect that the TTS diffusion process starts from structured data $\mathbf{X}_T \sim \mathcal{N}(\mathbf{E}_\theta(\mathbf{y}), \mathbf{I})$ rather than an unassuming Gaussian; the effect of these hyperparameters has not yet been explored for dysarthric TTS to the authors' knowledge. The reverse diffusion process

$$d\mathbf{X} = \frac{1}{2}((\mathbf{E}_\theta(\mathbf{y}) - \mathbf{X}) - \nabla \log p_t(\mathbf{X}))\beta(t)dt \quad (4)$$

is defined by a probability-flow ordinary differential equation (ODE) [33, 31]. Computing the log-density gradient $\nabla \log p_t(\mathbf{X})$ in (4), a.k.a. the *score*, is intractable, hence a trainable estimator $\mathbf{S}_\theta(\mathbf{X}, t, \mathbf{E}_\theta(\mathbf{y}), s_E)$ is necessary. \mathbf{S}_θ is a U-Net model [34] that must be trained on labelled speech data with a score-matching objective [33, 31]. When training on a dataset with multiple speakers, a speaker identity parameter s_E is given to control generation such that the output is faithful to the target speaker. The probability-flow ODE (4) is solved backwards in time from \mathbf{X}_T to \mathbf{X}_0 using the first-order Euler scheme. The HiFi-GAN vocoder [35] is then used to transform the mel-spectrograms \mathbf{X} into audio waveforms. This pipeline produces speaker-specific augmentation data for the subsequent ASR model to capture the nuances of dysarthric speakers. Although augmentation methods exist, which finetune a pretrained Grad-TTS model for VC [30], i.e. for a speech-to-speech task, we explore training Grad-TTS from scratch, i.e. for the more flexible TTS task and without reliance on matched control speech data nor typical TTS pre-training, providing augmentation that only requires labelled dysarthric data.

2.2. Model Adaptation for Dysarthric ASR

The data synthesised in Section 2.1 will be used to finetune the Whisper [12] ASR multilingual models.² Whisper is based on an encoder-decoder Transformer architecture with 12 encoder and 12 decoder layers and is a weakly supervised model trained using up to 680k hours of labelled typical speech data. To date, Transformer models have not been adequately explored in dysarthric ASR due to data sparsity issues, and adaptation of large-scale pre-trained ASR foundation models have not been previously explored. The Whisper model is finetuned using labelled data. Parameters in the feature encoder (2 x conv), model encoder and decoder layers were not frozen. The Whisper-medium (WM) model has 763.9M parameters (762.3M trainable parameters), and the Whisper-large (WL) model has 1543.3M parameters (1541.4M trainable parameters). Finally, SpecAugment [36] has been shown to improve the performance of dysarthric ASR with synthesised data augmentation [30], and therefore the Whisper models were also trained with and without SpecAugment in this work. SpecAugment is directly applied to the feature inputs of the ASR model, time warping the features, and masking blocks of frequency channels & blocks of time steps. Models with SpecAugment were optimised on the probability of frequency and time masking.

3. Experimental Setup

The *TORG*O database [37] containing dysarthric speech is introduced in Section 3.1, and the respective data splits & training methods for dysarthric speech synthesis and ASR models are described in Sections 3.2 and 3.3, respectively.

3.1. TORG O Dysarthric Speech Dataset

The *TORG*O database³ contains approx. 21 hours of aligned acoustic and 3D articulatory feature data [37], i.e. much less than usually used for ASR training. The data was gathered from 8 American English dysarthric speakers (with a diagnosis of amyotrophic lateral sclerosis (ALS), or cerebral palsy (CP)), denoted as *TORG*O dysarthric, and seven control speakers that are age-gender-matched to the dysarthric speakers, denoted as *TORG*O control. Utterances with no transcription or that were too short to contain speech were discarded [13]. All speakers had the same prompts, and therefore there is a large overlap in word and sentence prompts [38]. Since the *TORG*O dataset does not provide pre-defined data splits, a leave-one-speaker-out (LOSO) approach is commonly implemented for ASR [38]. The dysarthric speakers in *TORG*O were assessed by a speech and language therapist (SLT) using the Frenchay Dysarthria Assessment (FDA) [39]. The dysarthria severity ratings of the *TORG*O dysarthric speakers are displayed in Table 1. 'F' and 'M' denote gender, and the numeral denotes the participant number in the dataset.

Table 1: *Dysarthria severity for the TORG O database.*

	Severe	Mod.-Sev.	Moderate	Mild
Participant	F01, M01, M02, M04	M05	F03	F04, M03

²Whisper finetune code adapted from <https://github.com/vasistalodagala/whisper-finetune>.

³*TORG*O database: <http://www.cs.toronto.edu/~compilingweb/data/TORG O/torgo.html>

3.2. Text-to-Speech Synthesis

For dysarthric speech synthesis, we train the Grad-TTS models (cf. Section 2.1) using *TORGO dysarthric* from scratch. Previous implementations of DPM TTDS pre-train the synthesis model on age-gender-matched typical speech and finetune on dysarthric speech data [30, 40]. The Grad-TTS models require training and validation data for a given speaker to train a speaker embedding s_E in (1). *TORGO* does not have pre-defined data splits. Therefore, data splits were created for TTS training by pairing array and head microphones (of the same utterance) and then randomly splitting utterances into train, validation and test data splits in an 80%, 10%, 10% ratio per speaker. A systematic approach was considered, e.g. considering the distribution of single/multi-word utterances, and distribution of utterances across splits. However, analysis in [38] shows there are 951-969 unique utterances (across 16, 158 recordings), and not all dysarthric speakers completed recordings of all utterances. Once the *TORGO dysarthric* TTDS models are trained, the transcripts for all splits are input to the trained models to synthesise additional training data for LOSO [38] ASR model adaptation.

As Grad-TTS has not yet been adequately explored with dysarthric speech, we investigate β_T hyperparameter values in (3). Additionally, in the interest of using as little dysarthric data as possible, we investigate three conditions:

- (a) An *all-speaker (ASp) model* is trained using the data of all dysarthric speakers (i.e. the *TORGO dysarthric* data). The ASp model is trained using the training and validation splits of the entire dataset, and used to synthesise training data for all speakers.
- (b) *Single-speaker (SSp) models* are trained using a single dysarthric speaker’s data (i.e. a model is trained using one speaker’s training and validation data, and this model is used to synthesise data for the same speaker).
- (c) *Dysarthria-severity-group speaker (DSpG) models*: the *TORGO dysarthric* speakers are partitioned into two groups by dysarthria severity rating on the FDA (cf. Table 1). Severe and mod-severe speakers (i.e. F01, M01, M02, M04, M05) form Group 1 (G1), and mild and moderate speakers (i.e. F03, F04, M03) Group 2 (G2). The DSpG G1 model is trained on G1 speakers’ training and validation data, and used to synthesise data for these speakers.

Although control speaker data is not used to train models for dysarthric speech synthesis, ASp and SSp *TORGO control* TTS models were trained in the same manner to compare evaluation metrics. Finally, Grad-TTS models are optimised on hyperparameters of learning rate, epochs and batch size.

3.3. Dysarthric ASR

The Whisper ASR model is finetuned on a composition of real (*TORGO*) data and synthetic data (created by TTDS in Section 2.1). A cumulative ratio of additional synthetic training data for data augmentation (from 0-100% in 10% increments) is implemented. The *TORGO dysarthric* ASR models are trained using the LOSO methodology [38] to create speaker independent (SI) models: for a given target speaker, the data of the remaining speakers are used to train the model, which is tested on the given target speaker’s data (and therefore the target speaker’s data is not seen by the ASR model). Values for learning rate, warm-up, epochs, and batch size hyperparameters are optimised during training.

3.4. Evaluation metrics

To evaluate the quality of the synthesised dysarthric speech, the mean cepstral distortion (MCD) is used as an objective metric, and subjective evaluation by a human expert listener was conducted. The performance of dysarthric ASR systems are measured by WER.

3.4.1. Mean Cepstral Distortion (MCD)

The MCD is defined as the Euclidian distance between the synthesised and reference mel spectra, and is computed by alignment with dynamic time warping (DTW) [41]. The MCD has been shown to have correlation to subjective test results in speech synthesis analysis [42], although this has not been adequately investigated with dysarthric speech.

3.4.2. Subjective Evaluation

Subjective evaluation by expert listeners (SLTs) has been used to measure the presence and severity of dysarthric speech characteristics in synthesised speech [30]. For this study, an SLT with > 10 years of experience assessing and diagnosing speech disorders perceptually evaluated the synthesised data. For every dysarthric speaker, 20 audio samples from the *TORGO* database and 20 synthetic audio samples were randomly selected. The selected audio samples were presented to the SLT individually in random order. Every audio sample was rated on overall dysarthria severity on a 5-point scale (between 0 for none and 4 for severe) [1]. The mean scores were calculated for *TORGO* and synthetic audio per speaker, to allow comparison of the presence and severity of dysarthric speech characteristics.

3.4.3. Word Error Rate (WER)

Inference was performed with the pretrained baseline Whisper model, and adapted dysarthric Whisper models on a given target speaker (adapted models were trained on the remaining speakers in a LOSO methodology). Transcripts were processed with Whisper’s English text normalizer,⁴ and WER calculated between the processed reference and hypothesis transcripts. Average (Avg.) WER is calculated as the average of single-speaker WER scores, and severity group averages calculated as the average scores of speakers in the group (to allow direct comparison to similar studies [38]). The overall (Ovl.) WER score, commonly used to assess ASR for typical speech was also calculated by computing the WER score for transcripts across all speakers.

4. Results

4.1. Text to Speech Synthesis

MCD results for the *TORGO control* (C) and *TORGO dysarthric* (D) TTS models are shown in Table 2.

Table 2: MCD for GradTTS synthesis trained on *TORGO* data.

	C ASp	C SSp	D ASp	D DSpG	D SSp
$\beta_T = 10$	6.61	6.62	6.61	6.71	6.81
$\beta_T = 20$	6.75	7.92	6.98	7.49	7.80

Comparing MCD values for TTDS models with $\beta_T = 10$ and $\beta_T = 20$ in (3), the $\beta_T = 10$ models show lower MCD for

⁴Whisper normalizer: <https://github.com/openai/whisper/blob/main/whisper/normalizers>

control (C) and dysarthric (D) groups for all conditions ASp, SSp & DSpG (as defined in Section 3.2). Results indicate a slight tendency that more data leads to better (lower) MCD, however, for this the MCD difference is minor.

An SLT conducted a subjective evaluation (cf. Section 3.4.2) of TORGO data and data synthesised from the best model (i.e. D ASp), and the averaged ratings for dysarthria severity, their difference as well as the MCD metrics are displayed in Table 3. The ratings show that dysarthric speech characteristics are present in the synthesised samples, but there are differences in the level of severity. The Kendall’s Tau coefficient between average dysarthria severity scores and severity group MCD scores is -0.67 , indicating a strong negative association between the two ranked variables.

Table 3: Subjective evaluation of dysarthric data.

	Severe	Mod.-Sev.	Moderate	Mild
D severity ref.	3.28	2.45	1.70	0.275
D severity syn.	2.63	2.55	1.35	1.05
Difference	0.65	-0.10	0.35	-0.78
MCD	5.72	7.09	5.88	6.44

4.2. ASR model adaptation

4.2.1. Whisper baseline performance

The pretrained Whisper models (without any finetuning) are used for inference on the *TORGO control* (C) and *TORGO dysarthric* (D) data to establish baseline performance in the following. As a LOSO approach is used for *TORGO* ASR model adaptation, inference is performed on the whole dataset. Table 4 shows the performance in WER for the Whisper-medium (WM) and Whisper-large (WL) baseline models.

Table 4: WER in % for the Whisper medium (WM) and large (WL) baseline models. C denotes control and D dysarthric.

	Sev.	M.-S.	Mod.	Mild	Avg.	Ovl.
C WM	-	-	-	-	17.93	13.69
C WL	-	-	-	-	12.31	11.92
D WM	123.37	168.83	45.96	10.79	91.23	84.90
D WL	126.20	187.72	32.42	10.38	93.21	82.49

The WM and WL models for the *TORGO control* (non-dysarthric) data achieve overall WERs of 13.69 and 11.92%, respectively, and results for dysarthric speech shows much higher WER. The WL model has a relatively lower overall WER score on dysarthric speakers but a higher WER in average over speakers due to relatively poorer performance for severe and moderate-severe dysarthric speakers.

4.2.2. Whisper model adaptation

The Whisper medium (WM) and large (WL) models are finetuned on a composition of real data (*TORGO*) and an increasing percentage of synthetic data synthesised by the best TTDS model, i.e. D ASp with $\beta_T = 10$. Table 5 shows the results of the WM adaptation, since WM is the smaller model and showed better performance than WL for more severe dysarthric data. Adaptation using only real dysarthric data (i.e. no synthetic data) achieves an overall WER score of 56.14%, i.e. per-

Table 5: WER in % for *TORGO WM* model adaptation. Best performance in bold-face. * only 1 speaker.

Aug. %	Sev.	M.-Sev.*	Mod.*	Mild	Avg.	Ovl.
0	70.25	145.76	28.54	3.44	57.77	56.14
10	38.33	24.83	21.3	3.5	25.80	24.3
20	60.30	23.23	24.38	3.33	36.93	35.05
30	33.38	24.13	22.66	3.13	23.32	22.27
40	31.87	18.98	23.13	3.23	22.00	20.08
50	31.69	24.9	26.57	3.07	23.05	22.41
60	31.83	20.03	24.38	3.86	22.43	21.77
70	33.65	20.93	23.49	3.96	23.37	22.31
80	30.45	21.28	25.89	3.55	22.01	21.71
90	30.45	21.28	25.89	3.55	22.01	21.71
100	28.49	17.87	26.14	3.82	20.70	20.45

forms significantly better than the baseline in Table 4. Synthetic data further improves performance, with the best performance achieved using 100% additional synthesised training data, i.e. the training speakers in LOSO adaptation, reducing overall and average WER to 20.45% and 20.70%, respectively. Performance gains can be observed in particular for severe and moderate-severe dysarthric speech. The WL model also performed best with 100% additional synthesised data (not explicitly shown here), and the trend is maintained when the Whisper models are additionally trained with SpecAugment. Table 6 compares the WER performance of the WM and WL with 100% additional synthetic data (with and without SpecAugment) to recent SotA benchmarks on the *TORGO* ASR task. The proposed Grad-TTS augmented Whisper model adaptation outperforms all baseline models on the same task. The WL with SpecAugment shows best performance overall.

Table 6: *TORGO WER* performance in comparison to benchmarks. +denotes SpecAugment.

	Sev.	M.-Sev.	Mod	Mild	Avg.	Ovl.
LF-MMI [13]	-	-	-	-	42.90	-
FS2 & D-HMM [43]	55.88	49.60	36.80	12.60	39.20	-
FMLLR-DNN [44]	43.29	44.05	35.93	11.65	34.55	-
SD-CTL [45]	68.24	33.15	22.84	10.35	30.76	-
GTTS & WM	28.49	17.87	26.14	3.82	20.70	20.45
GTTS & WM+	25.82	14.95	5.26	4.28	19.16	18.35
GTTS & WL	46.39	18.29	24.27	3.03	25.84	25.28
GTTS & WL+	23.30	13.98	3.27	2.57	16.93	16.68

5. Conclusion

This work showed that it is possible to train Grad-TTS from scratch without matched control data to synthesise samples with dysarthric speech characteristics. A $\beta_T = 10$ schedule improves sample quality for the typical and dysarthric speech data used. The results show that Whisper can be finetuned for SotA DASR on *TORGO*, and that data augmentation is beneficial. The amount of synthesised data required is dependent on the severity of the dysarthric speaker.

6. References

- [1] J. R. Duffy, *Motor Speech Disorders*. Elsevier, 2019.
- [2] M. Walshe, *Clinical cases in dysarthria*. Milton Park, Abingdon, Oxon New York, NY: Routledge, 2022.
- [3] A. D. Page and K. M. Yorkston, “Communicative participation in dysarthria: Perspectives for management,” *Brain sciences*, vol. 12, no. 4, p. 420, 2022.

- [4] Z. C. Clarke, S. Judge, K. Fryer, S. Cunningham, J. Toogood, and M. S. Hawley, "A qualitative study exploring the effect of communicating with partially intelligible speech," *Augmentative and Alternative Communication*, vol. 39, no. 2, pp. 110–122, 2023.
- [5] H. Yusufali, R. K. Moore, and S. Goetze, "Refining Text Input for Augmentative and Alternative Communication (AAC) Devices: Analysing Language Model Layers for Optimisation," in *ICASSP 2024*, 2024.
- [6] D. J. Higginbotham, H. Shane, S. Russell, and K. Caves, "Access to AAC: Present, past, and future," *Augmentative and Alternative Communication*, vol. 23, no. 3, pp. 243–257, 2007.
- [7] S. Goetze, N. Moritz, J.-E. Appell, M. Meis, C. Bartsch, and J. Bitzer, "Acoustic user interfaces for ambient-assisted living technologies," *Informatics for Health and Social Care*, 2010.
- [8] N. Moritz, K. Adiloğlu, J. Anemüller, S. Goetze, and B. Kollmeier, "Multi-channel speech enhancement and amplitude modulation analysis for noise robust automatic speech recognition," *Computer Speech & Language*, vol. 46, 2017.
- [9] W. Zheng, A. Xiao, G. Keren, D. Le, F. Zhang, C. Fuegen, O. Kalinli, Y. Saraf, and A. Mohamed, "Scaling ASR Improves Zero and Few Shot Learning," in *Proc. Interspeech 2022*, 2022.
- [10] M. S. Hawley, P. Enderby, P. Green, S. Cunningham, S. Brownsell, J. Carmichael, M. Parker, A. Hatzis, P. O'Neill, and R. Palmer, "A speech-controlled environmental control system for people with severe dysarthria," *Medical Engineering & Physics*, vol. 29, no. 5, pp. 586–593, 2007.
- [11] F. Xiong, J. Barker, and H. Christensen, "Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition," in *ICASSP 2019*, 2019.
- [12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Int. Conf. on Machine Learning*, 2023.
- [13] E. Hermann and M. Magimai-Doss, "Dysarthric Speech Recognition with Lattice-Free MMI," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP20)*, 2020.
- [14] B. Vachhani, C. Bhat, B. Das, and S. K. Kopparapu, "Deep autoencoder based speech features for improved dysarthric speech recognition," in *Interspeech*, 2017, pp. 1854–1858.
- [15] C. Bhat and H. Strik, "Automatic assessment of sentence-level dysarthria intelligibility using blstm," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 322–330, 2020.
- [16] A. Almadhor, R. Irfan, J. Gao, N. Saleem, H. T. Rauf, and S. Kadry, "E2E-DASR: End-to-end deep learning-based dysarthric automatic speech recognition," *Expert Systems with Applications*, vol. 222, 2023.
- [17] S. R. Shahamiri, "Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system," *IEEE Trans. on Neural Systems and Rehabilitation Eng.*, vol. 29, 2021.
- [18] B. MacDonald, P.-P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, M. Ladewig, J. Tobin, M. Brenner, P. Q. Nelson *et al.*, "Disordered speech data collection: lessons learned at 1 million utterances from project euphonia," *Interspeech*, 2021.
- [19] A. Hernandez, P. A. Pérez-Toro, E. Nöth, J. R. Orozco-Arroyave, A. Maier, and S. H. Yang, "Cross-lingual self-supervised speech representations for improved dysarthric speech recognition," *arXiv preprint arXiv:2204.01670*, 2022.
- [20] L. P. Violeta, W.-C. Huang, and T. Toda, "Investigating self-supervised pretraining frameworks for pathological speech recognition," *arXiv preprint arXiv:2203.15431*, 2022.
- [21] Z. Chen, A. Rosenberg, Y. Zhang, G. Wang, B. Ramabhadran, and P. J. Moreno, "Improving Speech Recognition Using GAN-Based Speech Synthesis and Contrastive Unspoken Text Selection," in *Interspeech*, 2020.
- [22] Z. Jin, M. Geng, X. Xie, J. Yu, S. Liu, X. Liu, and H. Meng, "Adversarial data augmentation for disordered speech recognition," in *Proc. Interspeech 2021*, 2021, pp. 4803–4807.
- [23] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013.
- [24] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition," in *Interspeech*, 2018, pp. 471–475.
- [25] M. Geng, X. Xie, S. Liu, J. Yu, S. Hu, X. Liu, and H. Meng, "Investigation of data augmentation techniques for disordered speech recognition," *arXiv preprint arXiv:2201.05562*, 2022.
- [26] Z. Jin, M. Geng, J. Deng, T. Wang, S. Hu, G. Li, and X. Liu, "Personalized adversarial data augmentation for dysarthric and elderly speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.
- [27] W.-Z. Zheng, J.-Y. Han, C.-Y. Chen, Y.-J. Chang, and Y.-H. Lai, "Improving the efficiency of dysarthria voice conversion system based on data augmentation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, 2023.
- [28] M. Soleymanpour, M. T. Johnson, R. Soleymanpour, and J. Berry, "Synthesizing Dysarthric Speech using Multi-talker TTS for Dysarthric Speech Recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP22)*, 2022.
- [29] E. Hermann and M. Magimai-Doss, "Few-shot Dysarthric Speech Recognition with Text-to-Speech Data Augmentation," in *Proc. Interspeech*, 2023, pp. 156–160.
- [30] H. Wang, T. Thebaud, J. Villalba, M. Sydnor, B. Lammers, N. Dehak, and L. Moro-Velazquez, "DuTa-VC: A Duration-aware Typical-to-atypical Voice Conversion Approach with Diffusion Probabilistic Model," in *Proc. INTERSPEECH 2023*, 2023.
- [31] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *International Conference on Machine Learning*, 2021.
- [32] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [33] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Int. Conf. on Learning Representations*, 2021.
- [34] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.
- [35] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [36] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.
- [37] 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, Mar 2011.
- [38] Z. Yue, F. Xiong, H. Christensen, and J. Barker, "Exploring appropriate acoustic and language modelling choices for continuous dysarthric speech recognition," in *ICASSP 2020*, 2020.
- [39] P. Enderby, *Frenchay Dysarthria Assessment*. Pro-Ed, 1983.
- [40] H. Wang, V. Ravichandran, M. Rao, B. Lammers, M. Sydnor, N. Maragakis, A. A. Butala, J. Zhang, L. Clawson, V. Chovaz, and L. Moro-Velazquez, "Improving fairness for spoken language understanding in atypical speech with text-to-speech," in *Workshop on Synthetic Data Generation with Generative AI*, 2023.
- [41] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion," in *SLTU*, 2008, pp. 63–68.
- [42] A. N. Chadha, J. H. Nirmal, and P. Kachare, "A comparative performance of various speech analysis-synthesis techniques," *Int. J. of Signal Processing Systems*, vol. 2, no. 1, pp. 17–22, 2014.
- [43] M. Soleymanpour, M. T. Johnson, R. Soleymanpour, and J. Berry, "Accurate synthesis of dysarthric speech for asr data augmentation," *arXiv preprint arXiv:2308.08438*, 2023.
- [44] N. M. Joy and S. Umesh, "Improving acoustic models in torgo dysarthric speech database," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, pp. 637–645, 2018.
- [45] F. Xiong, J. Barker, Z. Yue, and H. Christensen, "Source domain data selection for improved transfer learning targeting dysarthric speech recognition," in *Proc. ICASSP20*, May 2020.