Proceedings Paper:
Leung, W.-Z. orcid.org/0009-0003-4888-1951, Christensen, H. and Goetze, S. (2025) Text-to-dysarthric-speech generation for dysarthric automatic speech recognition: is purely synthetic data enough? In: Speech and Computer: 27th International Conference, SPECOM 2025, Szeged, Hungary, October 13–15, 2025, Proceedings, Part I. SPECOM 2025, 13-15 Oct 2025, Szeged, Hungary. Lecture Notes in Computer Science (LNAI 16187). Springer Cham, pp. 203-216. ISBN: 9783032079558. ISSN: 0302-9743. EISSN: 1611-3349.

https://doi.org/10.1007/978-3-032-07956-5_14

Text-to-Dysarthric-Speech Generation for Dysarthric Automatic Speech Recognition: Is Purely Synthetic Data Enough?

Wing-Zin Leung[1 [0009−0003−4888−1951]], Heidi Christensen[1 [0000−0003−3028−5062]], and Stefan Goetze[1,2 [0000−0003−1044−7343]]

[1] Speech and Hearing (SPandH), School of Computer Science, The University of Sheffield, UK {wleung5, heidi.christensen}@sheffield.ac.uk
[2] South Westphalia University of Applied Sciences, Iserlohn, Germany goetze.stefan@fh-swf.de

Abstract. Recent advancements in text-to-speech (TTS) technology have revolutionised automatic speech recognition (ASR) data augmentation in low-resource settings. In particular, only a few public datasets are available for dysarthric ASR (DASR) and text-to-dysarthric-speech (TTDS) models have addressed data sparsity limitations by increasing training data samples and diversity. In this context, Grad-TTS (G-TTS) has been shown to synthesise speech with accurate dysarthric speech characteristics beneficial for DASR data augmentation; likewise, Matcha-TTS (M-TTS) has recently improved on typical speech synthesis baselines.

Recent studies commonly focus on data augmentation (i.e. reference data combined with additional synthetic data). This work analyses Whisper DASR model adaptation performance using reference data and G-TTS & M-TTS generated data, and shows that comparable performance can be achieved using synthesised data only relative to reference data. Additionally, despite growing work on dysarthric data augmentation, the validation of typical TTS metrics for synthetic dysarthric data, and the development of TTDS metrics requires further research. Results of this work show that gold standard metrics for typical TTS and current dysarthric speech assessment metrics lack sensitivity to predict DASR performance and hence a phoneme posteriorgram (PPG) distance based on the Jensen-Shannon divergence (JS) as a metric for dysarthric speech synthesis is introduced, showing correlation with downstream word error rate (WER) scores.

Keywords: Dysarthric speech recognition · Text-to-speech synthesis · Dysarthric TTS metrics.

1 Introduction

Dysarthria is a type of motor speech disorder (MSD) that reflects abnormalities in motor movements required for speech production [7]. The psychosocial impact

and restrictions on functioning and participation are well documented [5,50], and dysarthric ASR (DASR) has an important role in augmentative and alternative communication (AAC) devices and home control systems [16,11]. Although automatic speech recognition (ASR) systems achieved impressive performance with large-scale typical speech datasets, DASR performance is constrained by limited public availability of dysarthric data [3] and high inter- and intra-speaker variability inherent in dysarthric speech [38]. Challenges in data collection include recruitment and retention of participants with neurological conditions [33] and fatigue associated with MSDs [13], limiting collection of representative data and volume of data collected per target speaker. The TORGO database [37] is widely used in DASR studies, containing approx. 6 hours of acoustic data for 8 dysarthric speakers [18] - far less than typical speech datasets.

To improve ASR performance for low-resource dysarthric speech, studies have implemented various model adaptation methods [43,15,49], feature representations [1,17], and data selection methods [48]. Data augmentation techniques have been applied to increase training data, enhance diversity and mitigate overfitting [27,41], e.g. by vocal tract length perturbation (VTLP) [23], speed perturbation [10], or generative adversarial network (GAN)-based TTS [22]. Recent approaches aim to synthesise speech with accurate dysarthric speech characteristics (e.g. articulatory imprecision or voice quality [21]), and studies on dysarthric speech synthesis and DASR have focused on data augmentation (DAug), i.e. reference data in combination with additional ratios of synthetic data [27]. Recently, [41] leveraged a dysarthria level coefficient FastSpeech2 model to generate speech with accurate dysarthric features, showing improved word error rate (WER) when trained with additional synthetic data. Further, diffusion probabilistic modelling (DPM) is setting new standards across multiple domains and continuous-valued data generation tasks [6,31], and Grad-TTS (G-TTS) [35] trained on dysarthric data has been shown to synthesise speech with accurate dysarthric speech characteristics being beneficial for DASR DAug [27].

As the quality of synthetic dysarthric samples has improved, it is of interest to research DASR performance using only text-to-dysarthric-speech (TTDS) synthesised data and the potential of generating unseen speaker data to address e.g. an inter-variance gap. If out-of-domain speakers are generated, reference speaker data will not be available. Hence, this work is the first step towards this goal and explores diffusion probabilistic modelling TTDS synthesis and the utility of DASR model adaptation using purely synthetic data. Recent studies commonly focus on DAug, and this work addresses whether comparable performance can be achieved by using purely synthetic data only. The analysis shows comparable performance can be achieved for Whisper DASR model adaptation (Contribution 1 of this work): (i) using only TTDS synthesised data relative to reference data, and (ii) an equivalent volume of synthetic data can achieve comparable performance to reference data combined with additional synthetic data (i.e. DAug).

Finally, despite growing work on dysarthric data augmentation, research on metrics to evaluate the quality of synthesised dysarthric speech samples for

DASR is limited and typical speech synthesis metrics may not capture variation in dysarthric speech production, e.g. articulation [7]. Recent studies have focused on the synthesis of accurate dysarthric features by analysing (i) *subjective metrics*, e.g. ratings by clinicians on dysarthric features [27,46] and non-clinical listeners on naturalness [39] & similarity to dysarthric targets [41] and (ii) *objective metrics*, e.g. pitch contour [20] and intelligibility metrics [46] to show similarity or distance between reference and synthesised signals. Alternatively, studies have assessed TTDS quality by reporting on downstream DASR WER performance without evaluation metrics for generated samples [47,21]. However, here the final goal is to synthesise training data for improved DASR. Therefore, a reliable metric that correlates to DASR downstream performance will inform e.g. TTDS model & data selection for a target speaker, which is desirable to save time and computational resource in DASR system development. Furthermore, metrics designed for dysarthric speech are limited in number and validation. The pathological short-time objective intelligibility (P-STOI)/ pathological extended STOI (P-ESTOI) metrics [19] were originally designed to measure speech intelligibility for dysarthric speech signals, and studies have subsequently used the metrics for TTDS evaluation and shown correlation to rated severity in reference and synthetic samples [46]. However, overall dysarthria severity and communication-relevant parameters such as intelligibility [24] do not necessarily reflect that dysarthric pathomechanisms underlying e.g. intelligibility impairment are captured, and these measures have not yet been validated as metrics to predict downstream DASR performance. Therefore, in this work (Contribution 2): (i) synthesis metrics for TTDS data and downstream DASR performance are evaluated, (ii) a PPG distance metric is introduced as a measure of pronunciation distance for generated dysarthric speech to capture similarity in articulation[3] and (iii) correlation analysis shows that current metrics lack sensitivity to predict DASR performance for synthetic dysarthric data, and that a lower PPG distance is associated with lower WER scores.

2 Experimental Setup

The *TORGO* dysarthric speech database [37] as data source is briefly described in Section 2.1. Section 2.2 introduces the TTDS models, including training methods and model evaluation. Section 2.3 introduces Whisper and DASR model adaptation, and adaptation experiments. Finally, the evaluation metrics for TTDS and DASR, and correlation analysis for metrics and downstream DASR performance are introduced in Section 2.4.

2.1 The TORGO Dysarthric Speech Dataset

The *TORGO* database contains approx. 15 hours of acoustic data [37], far less than usually used for ASR model training. Data in the *TORGO* dataset

[3] A PPG is a time-varying categorical distribution over speech units (e.g. phonemes) [14] and recent work has demonstrated interpretable pronunciation distance [4].

was gathered from 8 dysarthric speakers with a diagnosis of cerebral palsy or amyotrophic lateral sclerosis (denoted as *TORGO* dysarthric (TD)), and 7 age-gender-matched control speakers (denoted as *TORGO* control (TC)). Non-speech utterances and utterances with no transcription were discarded [15], and utterances with direct instruction were corrected (e.g. ['Lead' as in 'I will lead you'] to ['Lead']) [27]. Manual listening was conducted to determine audio length filtering, identifying samples that are too short to contain speech (< 0.4 seconds), and samples that incorrectly contain multiple utterances (> 60 seconds). The dysarthric speakers in *TORGO* were assessed by a speech and language therapist (SLT) using the Frenchay Dysarthria assessment (FDA) [8]. The severity ratings are: *severe* for speakers F01, M01, M02 & M04, *moderate-severe* for speaker M05, *moderate* for speaker F03, and *mild* for speakers F04 & M03 [37]. 'F' and 'M' denote gender, and the numeral denotes the participant number in the dataset.

2.2 Text-to-Dysarthric Speech (TTDS) Synthesis Models

The Grad-TTS (G-TTS) and Matcha-TTS (M-TTS) models are selected for TTDS synthesis in this work. G-TTS trained on dysarthric data has been shown to generate samples with accurate dysarthric speech characteristics that are beneficial for DASR [27], and M-TTS shows improved objective and subjective performance relative to DPM TTS baselines [30]. In G-TTS, mel spectrograms are generated with a score-based decoder (defined by a probability flow ordinary differential equation (ODE) [42]) from monotonic alignment search-aligned encoder outputs [35]. M-TTS [30] introduces innovation to non-autoregressive TTS with optimal-transport conditional flow-matching [29] to learn ODEs that sample from a data distribution. A HiFi-GAN [26] vocoder (trained on the LibriTTS dataset [51])[4] is used to transform mel spectrograms generated by TTS models into audio waveforms.

The models are trained from scratch using *TORGO* data to create Grad-TTDS (G-TTDS)[5] and Matcha-TTDS (M-TTDS)[6] models, respectively. As both models require training and validation data for a given speaker in order to train a speaker embedding, data splits as in [27] were created for TTDS model training by pairing array and microphone recordings (of the same utterance), and then randomly splitting utterances into train, validation, and test splits in an 80%, 10%, and 10% ratio per speaker, respectively. TTDS models are evaluated on only the test split, and metrics are calculated between the reference *TORGO* dysarthric (TD) data and equivalent TTDS synthesised data (i.e. generated using the equivalent text label and corresponding speaker embedding). The transcripts

[4] HiFIGAN LibriTTS 16kHz vocoder: `https://huggingface.co/speechbrain/tts-hifigan-libritts-16kHz`.

[5] G-TTS training code adapted from `https://github.com/WingZLeung/TTDS`.

[6] M-TTS training code adapted from `https://github.com/shivammehta25/Matcha-TTS`. Code & audio samples available at `https://github.com/WingZLeung/M-TTDS`.

for all data splits are input to the trained TTDS models to synthesise a complete TD dataset for DASR model adaptation.

2.3 Whisper DASR Model Adaptation

The *TORGO* data (cf. Section 2.1) and synthesised data (cf. Section 2.2) are used to finetune 3 Whisper [36] ASR multilingual models[7] with encoder-decoder Transformer architecture, originally trained weakly supervised on 680k hours of typical speech. The Whisper-medium (WM) model has 24 encoder and decoder layers and 769M parameters, the Whisper-large (WL) model has 32 encoder and decoder layers and 1550M parameters and the Whisper-large-v2 (WL2) model is trained for 2.5x more epochs with SpecAugment [34] and added regularisation [36]. The Whisper models are fine-tuned using labelled data. Parameters in the feature encoder (2x conv.), model encoder and decoder layers are not frozen.

The leave-one-speaker-out (LOSO) evaluation methodology is used for DASR model adaptation to be consistent with [9] (and subsequent work, e.g. [15,49]) to create speaker-independent models. Hyperparameters for learning rate, warm-up, epochs, and batch size are optimised during model adaptation via grid search with the best checkpoint selected by the lowest validation WER. To investigate the performance of the DASR model adaptation using purely synthetic data relative to reference data, 2 experiments are conducted. For Experiment 1, DASR model adaptation performance using either reference data, or G-TTDS or M-TTDS synthesised data is compared. As recent studies commonly focus on data augmentation (DAug) (i.e. reference dysarthric data in combination with TTDS data) to increase training data and enhance sample diversity, for Experiment 2, DAug (1:1 ratio reference:synthetic data) is compared to an equivalent number of G-TTDS and M-TTDS combined samples (1:1 ratio G-TTDS:M-TTDS data, i.e. same proportion) to compare an equivalent volume of synthetic data with sample diversity from 2 TTDS models. Studies commonly use both the *TORGO* dysarthric and control data for DASR model adaptation. To reduce computation, only dysarthric data is used for DASR in this study.

2.4 Evaluation Metrics for TTDS and DASR

Despite growing interest in dysarthric data augmentation, research on validating typical speech TTS metrics for dysarthric speech, and the development of metrics designed for TTDS is limited. Therefore, gold standard typical speech metrics as well as metrics for dysarthric speech assessment are investigated. The metrics are computed to evaluate TTDS models (cf. Section 2.2), and the performance of DASR systems are measured by WER. To evaluate the utility of metrics for TTDS data to indicate downstream DASR performance, correlation analysis between metrics and DASR system WER performance is conducted. The metrics used in this work are briefly described below:

[7] Whisper finetune code adapted from `https://github.com/vasistalodagala/whisper-finetune`.

MCD: The mel cepstral distortion (MCD) is defined as the Euclidian distance between a reference mel spectrum and time aligned synthesised spectrum, and is computed by alignment with dynamic time warping (DTW) [25]. The MCD has been shown to have correlation to subjective listening test results in TTS [2] and TTDS [27] model performance.

L-f_0: Log f_0 root mean square error (RMSE) refers to the logarithmic fundamental frequency (f_0) contour RMSE of a reference signal compared to the respective logarithmic f_0 contour of a synthesised signal [45]. DTW is computed for alignment, and the metric calculation is based only on voiced frames of the speech signal.

P-STOI/P-ESTOI: The pathological short-time objective intelligibility (P-STOI)/pathological extended STOI (P-ESTOI) metrics [19] are designed to measure speech intelligibility (secondary to motor speech production deficits), by quantifying distortion in time-frequency structure between control and dysarthric speech signals [19]. Studies have shown correlation to dysarthria severity in reference and synthetic samples [46]. The short-time correlation or spectral correlation between one-third octave band representations of reference and time-aligned test signal yields the P-STOI and P-ESTOI metrics, respectively. As in [19], octave band representation alignment is achieved by DTW (using the Euclidean distance as the cost function).

PPG-D: The phoneme posteriorgram (PPG) is a time-varying categorical distribution over acoustic speech units, e.g. phonemes [14], and studies have shown effective application to downstream dysarthric speech tasks, including voice conversion [52] and classification [12]. The High-fidelity Neural (H-FN) PPG model [4] has been shown to encode interpretable pronunciation distance, and therefore a metric using PPGs to measure pronunciation error for generated dysarthric speech is investigated. Inference is performed using the H-FN PPG model[8] to output PPGs of dimension (phonemes, frames). The Jensen-Shannon divergence (JS) [28]

$$L(P,Q) = H(aP + (1-a)Q) - aH(P) - (1-a)H(Q) \tag{1}$$

is calculated between a reference posteriorgram PPG(P) and DTW time-aligned test posteriorgram PPG(Q) to compute the PPG distance (PPG-D). In (1), a, $0 < a < 1$ weights the two PPG probability distributions over M frames of $P = (p_1, ..., p_M)$ and $Q = (q_1, ..., q_M)$, respectively, and $H(P) = -\Sigma_{i=1}^{M} p_i \log p_i$ is Shannon's entropy [40].

WER: ASR transcripts are pre-processed with Whisper's English text normalizer[9] before word error rate (WER) is calculated between processed hypothesis and reference (ground-truth) transcripts. Both, (i) overall (Ovl.) and (ii) average (Avg.) WER scores are calculated [27] by (i) computing the WER scores for transcripts across all speakers and (ii) average WER scores of single-speaker WERs or average severity group WERs.

[8] H-FN PPG model: `https://github.com/interactiveaudiolab/ppgs`.

[9] Whisper normalizer: `https://github.com/openai/whisper/blob/main/whisper/normalizers`.

Correlation analysis: A Spearman's rank-order correlation analysis is conducted to assess the monotonic relationship between metrics and DASR performance (i.e. WER) for non-parametric data. Intrusive metrics are calculated between the reference TD data and equivalent TTDS synthesised data (i.e. generated using the equivalent text label and corresponding speaker embedding), and correlated to the WER score for the given TD utterance. For the analysis, transcripts that are common between all dysarthric speakers are selected to allow comparison of metrics between equivalent transcripts.

3 Results

3.1 Text-to-Dysarthric-Speech Synthesis model evaluation

The G-TTDS and M-TTDS models are trained from scratch on the TD data. Evaluation metrics are computed on the test set created for TTDS model training (cf. Section 2.2). Table 1 shows the results of the objective intrusive metrics (i.e. calculated between the reference TD data and equivalent TTDS synthesised data).

Table 1. TTDS model evaluation.

	MCD ↓	L-f_0 ↓	PSTOI ↑	PESTOI ↑	PPG-D ↓
G-TTDS	**6.59**	0.40	0.37	0.23	0.64
M-TTDS	7.47	**0.38**	**0.40**	**0.26**	**0.56**

The M-TTDS model achieves better scores for all metrics apart from MCD, indicating that M-TTDS data is more similar in f_0 pitch contour and estimated speech intelligibility to reference *TORGO* dysarthric data than G-TTDS data. Informal listening tests by an SLT further showed that M-TTDS data is more similar to reference data in dysarthria severity level and accuracy of dysarthric speech characteristics. Thus, TTDS model evaluation indicates higher quality synthesis and similarity to reference data for M-TTDS samples. The following DASR model adaptation will investigate whether samples from a model with higher evaluation metrics will have better downstream DASR performance, and analyse the correlation between metrics and WER performance.

3.2 Pre-trained Whisper Model Baseline Inference

The pretrained Whisper models (i.e. without any finetuning) are first used for inference on the TD data, and synthesised data from the G-TTDS or M-TTDS models to establish baseline performance for ASR systems trained on typical data. Inference is performed on the whole TD dataset (and on the dataset generated by the TTDS models from TD transcripts). Table 2 shows the average (Avg.), overall (Ovl.) and per-severity-group WER for the pretrained Whisper models.

Table 2. WER in % for the pretrained WM, WL and WL2 models on the *TORGO* dysarthric (TD) data, and G-TTDS and M-TTDS synthesised data.

Model	Data	Severe	M.-S.	Mild	Avg.	Ovl.
WM	TD	115.90	152.40	20.27	84.60	77.97
WL	TD	127.08	186.55	17.37	93.37	82.30
WL2	TD	96.21	93.80	20.79	67.63	63.01
WM	G-TTDS	157.17	235.89	70.20	134.40	128.90
WL	G-TTDS	148.97	172.20	66.43	120.92	116.74
WL2	G-TTDS	145.24	112.61	66.50	111.64	105.95
WM	M-TTDS	115.50	40.07	38.90	77.35	75.34
WL	M-TTDS	96.63	56.66	38.63	69.88	68.68
WL2	M-TTDS	92.08	45.51	33.16	64.16	62.11

As expected, all ASR models show high WER on the TD data, in particular for *severe* and *moderate to severe (M.-S.)* dysarthric speech, with relatively better performance for the WL2 model. For G-TTDS data, all pre-trained Whisper models show significantly higher average WER scores relative to the TD data, and perform worse for all speakers. For M-TTDS data, all pre-trained Whisper models achieve better average WER performance relative to the TD data (by 7.25%, 23.49% and 3.47% for the WM, WL, and WL2 models, respectively). Comparing severity groups, the M-TTDS data leads to marginally better performance for *severe* speakers, significantly better performance for *M.-S.* speakers, and worse performance for *mild* speakers overall. Results in Table 2 are in line with performance metrics in Table 1 as well as other work demonstrating correlation between WER and speech intelligibility for dysarthric speech [44].

3.3 Whisper DASR model adaptation performance

Experiment 1: DASR model adaptation using either reference or synthetic data only Whisper models are adapted with a LOSO methodology using either the speaker-specific TD training data, or the equivalent speaker-specific training data synthesised by the G-TTDS or M-TTDS models. All models are tested on the TD data (i.e. reference audio data) for the given target speaker.

Table 3 (top table, Experiment 1) shows the performance in WER for adapted Whisper models. In general, model adaptation significantly improves performance relative to the baselines reported in Table 2. Comparing adaptation with real (TD) vs. synthesized data, G-TTDS data shows better WER performance for the WM & WL2 models (by 3.55% & 14.42% average WER, respectively), but a higher average WER for the WL model (by 7.45%) while M-TTDS shows better WER performance for the WL & WL2 models (by 3.11% & 12.45% average WER, respectively), and higher WER for the WM model (by 5.12% average WER). The best performing model overall is the M-TTDS WL model, which in particular shows the best performance on severe speakers. Therefore, the best

Table 3. WER in % for adapted Whisper Models. Experiment 1 and Experiment 2 results.

Experiment 1

Model	Data	Sev.	M.-S.	Mild	Avg.	Ovl.
WM	TD	65.93	44.39	18.62	45.50	41.95
WL	TD	43.78	21.11	15.39	30.30	28.83
WL2	TD	71.70	**19.16**	12.20	42.82	38.11
WM	G-TTDS	62.86	40.91	14.42	41.95	35.60
WL	G-TTDS	59.59	29.55	11.35	37.75	30.36
WL2	G-TTDS	42.74	24.88	10.46	28.40	27.78
WM	M-TTDS	85.62	27.46	11.67	50.62	46.57
WL	M-TTDS	**41.08**	24.18	**9.67**	**27.19**	**25.34**
WL2	M-TTDS	44.48	25.64	13.13	30.37	29.19

Experiment 2

Model	Data	Sev.	M.-S.	Mild	Avg.	Ovl.
WM	TD+G-TTDS	48.72	20.35	12.41	31.56	31.93
WL	TD+G-TTDS	31.99	**17.14**	8.30	21.25	19.85
WL2	TD+G-TTDS	**29.41**	19.51	8.10	20.18	**18.37**
WM	TD+M-TTDS	54.11	32.47	14.95	36.72	34.51
WL	TD+M-TTDS	41.45	18.05	8.41	26.14	24.86
WL2	TD+M-TTDS	29.94	17.63	**7.58**	**20.02**	18.86
WM	G-TTDS+M-TTDS	48.87	35.96	15.97	34.92	30.39
WL	G-TTDS+M-TTDS	32.96	22.51	9.92	23.02	21.82
WL2	G-TTDS+M-TTDS	35.06	24.04	10.87	24.61	21.54

performing WM, WL, & WL2 models only use TTDS data, and synthetic data only outperformed reference data.

Notably, although the M-TTDS model shows the best evaluation metrics (cf. Table 1), adaptation using M-TTDS data does not consistently achieve the best WER performance, highlighting that (i) synthetic data that is more similar to reference data is not necessarily better for DASR model adaptation and (ii) higher quality and similarity to reference data as measured by current TTDS model evaluation is not sufficient to inform downstream DASR performance. To determine if metrics are able to provide an indication of downstream DASR performance, Spearman's correlation (ρ) is calculated between metrics and WER performance.

Table 4 shows the Spearman's correlation (ρ) between TTS metrics and WER performance. There is a weak to moderate relationship between PPG-D and WER for all models ($P< 0.001$), and a negligible to low correlation between MCD, L-f_0, P-STOI, P-ESTOI and WER. Comparing data, M-TTDS models have higher PPG-D ρ values relative to G-TTDS models, but do not consistently have higher WERs (Kendall's Tau coefficient between PPG-D ρ and WER across all models and data = 0.6). In summary, there is a trend between a higher

Table 4. TTS metrics and WER correlation (Experiment 1). Values = Spearman's correlation (ρ).

Model	Data	MCD	L-f_0	PSTOI	PESTOI	PPG-D
WM	G-TTDS	0.045	0.153	-0.035	0.014	0.441
WL	G-TTDS	0.035	0.095	-0.027	-0.019	0.284
WL2	G-TTDS	0.009	0.065	-0.034	-0.002	0.252
WM	M-TTDS	-0.033	0.117	-0.001	0.048	0.497
WL	M-TTDS	0.023	0.079	-0.041	-0.010	0.360
WL2	M-TTDS	0.004	0.099	-0.022	0.031	0.423

PPG-D score and higher WER score across all models, which is not observed for other metrics. Therefore, PPG-D can provide an indication of downstream DASR performance as an evaluation metric for TTDS data, while current TTS and dysarthric speech assessment metrics lack sensitivity.

Experiment 2: DASR model adaptation and data augmentation (DAug) Whisper models are again adapted with a LOSO methodology to compare adaptation with (i) DAug and (ii) purely synthetic data. For (i): TD data is used in combination with either G-TTDS or M-TTDS synthesised data (1:1 ratio) and for (ii): combined G-TTDS+M-TTDS data (1:1 ratio) is used to compare adaptation with an equivalent number of samples (as (i)) with sample diversity from both TTDS models. Table 3 (bottom table, Experiment 2) shows the performance in WER for these adapted Whisper models.

In general, WER performance is improved further relative to only using TD, G-TTDS or M-TTDS data independently (cf. Table 3, top). The TD+G-TTDS data shows the best WM and WL model average WER performance, and marginally worse performance than TD+M-TTDS WL2 (by 0.16% average WER), and achieves the best average WER scores for *severe* and *moderate-severe* speakers. The G-TTDS+M-TTDS models show better WER performance than the TD+M-TTDS WM and WL models (by 1.8% & 3.12% average WER, respectively). The WER performance for synthesised data only is comparable to DAug, although TD+G-TTDS data achieve the best performance by marginal scores.

Table 5 shows the Spearman's correlation (ρ) between TTS metrics and WER performance for Experiment 2. The Spearman's correlation (ρ) between metrics and WER performance for Experiment 2 are similar to results for Experiment 1 - there is weak to moderate correlation between PPG-D and WER for all models ($P< 0.001$), and negligible to low correlation observed for other metrics. Therefore, PPG-D can also provide an indication of downstream DASR performance for DAug, while other metrics lack sensitivity. Although TTDS model evaluation shows that G-TTDS data is less similar to reference TD data (cf Section 3.1), G-TTDS data shows the overall best performance for DAug. Studies have shown that enhancing diversity with DAug is beneficial for pathological ASR [32], and the G-TTDS data also has a higher range (R) and standard deviation (SD) of

Table 5. TTS metrics and WER correlation (Experiment 2).

Model	Data	MCD	L-f_0	PSTOI	PESTOI	PPG-D
WM	TD+G-TTDS	0.055	0.111	-0.040	0.015	0.377
WL	TD+G-TTDS	0.022	0.070	-0.022	-0.017	0.234
WL2	TD+G-TTDS	0.030	0.069	-0.041	-0.027	0.214
WM	TD+M-TTDS	-0.015	0.096	-0.027	0.025	0.495
WL	TD+M-TTDS	0.028	0.057	-0.057	-0.031	0.285
WL2	TD+M-TTDS	0.049	0.055	-0.066	-0.041	0.284
WM	G-TTDS+M-TTDS	0.019	0.052	-0.041	0.001	0.279
WL	G-TTDS+M-TTDS	0.024	0.056	-0.025	-0.022	0.213
WL2	G-TTDS+M-TTDS	-0.003	0.067	-0.032	-0.018	0.229

PPG-D values relative to M-TTDS data, particularly for *severe* to *M-S.* speakers (G-TTDS: Avg.=0.732, R=2.16, SD=0.32, M-TTDS: Avg.=0.645, R=1.92, SD=0.28). Therefore, a degree of distance in similarity to reference data and increased sample diversity seems to be beneficial for DAug.

4 Conclusion

This work shows that generative TTDS models can successfully create augmentation data for DASR. In particular, that it is possible and even beneficial to use synthetic data only (relative to using reference data) for Whisper model adaptation, potentially due to a more diverse data distribution relative to small datasets. Performance is improved further with DAug (i.e. reference data in combination with additional synthetic data), and synthetic data only (i.e. an equivalent volume of synthetic data from 2 TTDS models) has comparable performance, indicating that more data with more diversity is important for data augmentation when using reference or synthetic data. The analysis further highlights that most current TTS metrics lack sensitivity to predict DASR performance, while a trend between PPG-D and WER score is shown.

Acknowledgments. 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].

References

1. Cadet, X.F., Aloufi, R., Ahmadi-Abhari, S., Haddadi, H.: A study on the impact of self-supervised learning on automatic dysarthric speech assessment. In: ICASSP Workshops (ICASSPW) (2024)
2. Chadha, A.N., Nirmal, J.H., Kachare, P.: A comparative performance of various speech analysis-synthesis techniques. Int. J. of Signal Processing Systems **2**(1), 17–22 (2014)

3. Christensen, H., Casanueva, I., Cunningham, S., Green, P., Hain, T.: Automatic selection of speakers for improved acoustic modelling: Recognition of disordered speech with sparse data. In: 2014 IEEE Spoken Language Technology Workshop (SLT). pp. 254–259. IEEE (2014)
4. Churchwell, C., Morrison, M., Pardo, B.: High-fidelity neural phonetic posteriorgrams. arXiv preprint arXiv:2402.17735 (2024)
5. Clarke, Z.C., Judge, S., Fryer, K., Cunningham, S., Toogood, J., Hawley, M.S.: A qualitative study exploring the effect of communicating with partially intelligible speech. Augmentative and Alternative Communication **39**(2), 110–122 (2023)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
7. Duffy, J.R.: Motor Speech Disorders. Elsevier (2019)
8. Enderby, P.: Frenchay Dysarthria Assessment. Pro-Ed (1983)
9. Espana-Bonet, C., Fonollosa, J.A.: Automatic speech recognition with deep neural networks for impaired speech. In: IberSPEECH (2016)
10. Geng, M., Xie, X., Liu, S., Yu, J., Hu, S., Liu, X., Meng, H.: Investigation of data augmentation techniques for disordered speech recognition. arXiv preprint arXiv:2201.05562 (2022)
11. Goetze, S., Moritz, N., Appell, J.E., Meis, M., Bartsch, C., Bitzer, J.: Acoustic user interfaces for ambient-assisted living technologies. Informatics for Health and Social Care (2010). https://doi.org/10.3109/17538157.2010.528655
12. Gosztolya, G., Svindt, V., Bóna, J., Hoffmann, I.: Extracting phonetic posteriorbased features for detecting multiple sclerosis from speech. IEEE Trans. on Neural Systems and Rehabilitation Engineering (2023)
13. Hartelius, L., Burge, Å., Johansson, A., Ljungsfors, A., Mattsson, A., Winkworth, A., Andersen, O.: How does fatigue affect communication? the influence of fatigue on cognitive, physical, psychosocial and communicative ability in individuals with multiple sclerosis. International Journal of MS Care **6**(2), 39–51 (2004)
14. Hazen, T.J., Shen, W., White, C.: Query-by-example spoken term detection using phonetic posteriorgram templates. In: ASRU (2009)
15. Hermann, E., Magimai.-Doss, M.: Dysarthric Speech Recognition with Lattice-Free MMI. In: ICASSP (2020). https://doi.org/10.1109/ICASSP40776.2020.9053549
16. Higginbotham, D.J., Shane, H., Russell, S., Caves, K.: Access to AAC: Present, past, and future. Augmentative and Alternative Communication **23**(3), 243–257 (2007). https://doi.org/10.1080/07434610701571058
17. Hu, S., Xie, X., Geng, M., Jin, Z., Deng, J., Li, G., Wang, Y., Cui, M., Wang, T., Meng, H., et al.: Self-supervised ASR models and features for dysarthric and elderly speech recognition. IEEE/ACM Trans. on Audio, Speech, and Language Processing (2024)
18. Hui, M., Zhang, J., Mohan, A.: Enhancing aac software for dysarthric speakers in e-health settings: An evaluation using torgo. arXiv preprint arXiv:2411.00980 (2024)
19. Janbakhshi, P., Kodrasi, I., Bourlard, H.: Pathological Speech Intelligibility Assessment Based on the Short-time Objective Intelligibility Measure. In: ICASSP (2019). https://doi.org/10.1109/ICASSP.2019.8683741
20. Jiao, Y., Tu, M., Berisha, V., Liss, J.: Simulating dysarthric speech for training data augmentation in clinical speech applications. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) (2018)

21. Jin, Z., Geng, M., Deng, J., Wang, T., Hu, S., Li, G., Liu, X.: Personalized adversarial data augmentation for dysarthric and elderly speech recognition. IEEE/ACM Trans. on Audio, Speech, and Language Processing (2023)
22. Jin, Z., Geng, M., Xie, X., Yu, J., Liu, S., Liu, X., Meng, H.: Adversarial data augmentation for disordered speech recognition. In: Interspeech. pp. 4803–4807 (2021). https://doi.org/10.21437/Interspeech.2021-168
23. Kanda, N., Takeda, R., Obuchi, Y.: Elastic spectral distortion for low resource speech recognition with deep neural networks. In: ASRU (2013)
24. Klopfenstein, M., Bernard, K., Heyman, C.: The study of speech naturalness in communication disorders: A systematic review of the literature. Clinical linguistics & phonetics **34**(4) (2020)
25. Kominek, J., Schultz, T., Black, A.W.: Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In: SLTU. pp. 63–68 (2008)
26. Kong, J., Kim, J., Bae, J.: HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. Advances in Neural Information Processing Systems **33** (2020)
27. Leung, W.Z., Cross, M., Ragni, A., Goetze, S.: Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis. In: Interspeech. Kos, Greece (Sep 2024). https://doi.org/10.21437/Interspeech.2024-1645
28. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory **37**(1), 145–151 (1991)
29. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
30. Mehta, S., Tu, R., Beskow, J., Székely, É., Henter, G.E.: Matcha-TTS: A fast TTS architecture with conditional flow matching. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) (2024)
31. Mehta, S., Wang, S., Alexanderson, S., Beskow, J., Székely, É., Henter, G.E.: Diff-ttsg: Denoising probabilistic integrated speech and gesture synthesis. arXiv preprint arXiv:2306.09417 (2023)
32. Mujtaba, D., Mahapatra, N.R., Arney, M., Yaruss, J.S., Herring, C., Bin, J.: Inclusive asr for disfluent speech: Cascaded large-scale self-supervised learning with targeted fine-tuning and data augmentation. arXiv preprint arXiv:2406.10177 (2024)
33. Newberry, A., Sherwood, P., Hricik, A., Bradley, S., Kuo, J., Crago, E., Hoffman, L.A., Given, B.A.: Understanding recruitment and retention in neurological research. J. Neuroscience Nursing **42**(1) (2010)
34. Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D., Le, Q.V.: Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779 (2019)
35. Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., Kudinov, M.: Grad-tts: A diffusion probabilistic model for text-to-speech. In: International Conference on Machine Learning (2021)
36. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: Int. Conf. on Machine Learning (2023)
37. Rudzicz, F., Namasivayam, A.K., Wolff, T.: The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Language Resources and Evaluation **46**(4), 523–541 (Mar 2011). https://doi.org/10.1007/s10579-011-9145-0
38. Rudzicz, F.: Using articulatory likelihoods in the recognition of dysarthric speech. Speech Communication **54**(3), 430–444 (2012)

39. Shahamiri, S.R.: Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system. IEEE Transactions on Neural Systems and Rehabilitation Engineering **29** (2021)
40. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948)
41. Soleymanpour, M., Johnson, M.T., Soleymanpour, R., Berry, J.: Accurate synthesis of dysarthric speech for asr data augmentation. Speech Communication **164**, 103112 (2024)
42. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: Int. Conf. on Learning Representations (2021)
43. Takashima, R., Takiguchi, T., Ariki, Y.: Two-step acoustic model adaptation for dysarthric speech recognition. In: ICASSP (2020)
44. Van Nuffelen, G., Middag, C., De Bodt, M., Martens, J.P.: Speech technology-based assessment of phoneme intelligibility in dysarthria. International journal of language & communication disorders (2009)
45. Wang, C.C., Ling, Z.H., Zhang, B.F., Dai, L.R.: Multi-layer F0 modeling for HMM-based speech synthesis. In: 2008 6th International symposium on Chinese spoken language processing. IEEE (2008)
46. Wang, H., Thebaud, T., Villalba, J., Sydnor, M., Lammers, B., Dehak, N., Moro-Velazquez, L.: Duta-vc: A duration-aware typical-to-atypical voice conversion approach with diffusion probabilistic model. arXiv preprint arXiv:2306.10588 (2023)
47. Wang, H., Jin, Z., Geng, M., Hu, S., Li, G., Wang, T., Xu, H., Liu, X.: Enhancing pre-trained ASR system fine-tuning for dysarthric speech recognition using adversarial data augmentation. In: ICASSP (2024)
48. Xiong, F., Barker, J., Yue, Z., Christensen, H.: Source domain data selection for improved transfer learning targeting dysarthric speech recognition. In: ICASSP (2020)
49. Yue, Z., Xiong, F., Christensen, H., Barker, J.: Exploring appropriate acoustic and language modelling choices for continuous dysarthric speech recognition. In: ICASSP 2020 (2020). https://doi.org/10.1109/icassp40776.2020.9054343
50. Yusufali, H., Moore, R.K., Goetze, S.: Refining Text Input for Augmentative and Alternative Communication (AAC) Devices: Analysing Language Model Layers for Optimisation. In: ICASSP 2024 (2024)
51. Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R.J., Jia, Y., Chen, Z., Wu, Y.: Libritts: A corpus derived from librispeech for text-to-speech. arXiv preprint arXiv:1904.02882 (2019)
52. Zheng, W.Z., Han, J.Y., Cheng, H.L., Chu, W.C., Chen, K.C., Lai, Y.H.: Comparing the performance of classic voice-driven assistive systems for dysarthric speech. Biomedical Sig. Proc. and Control (2023)