Improving fairness for spoken language understanding in atypical speech with Text-to-Speech

Helin Wang
Johns Hopkins University
hwang258@jhu.edu

Venkatesh Ravichandran
Amazon
veravic@amazon.com

Milind Rao
Amazon
milinrao@amazon.com

Becky Lammers
Johns Hopkins University School of Medicine
blammer2@jh.edu

Myra J. Sydnor
Johns Hopkins University School of Medicine
msydnor3@jhmi.edu

Nicholas Maragakis
Johns Hopkins University School of Medicine
nmaragak@jhmi.edu

Ankur A. Butala
Johns Hopkins University School of Medicine
Ankur.Butala@jhmi.edu

Jayne Zhang
Johns Hopkins University School of Medicine
jz@jhmi.edu

Lora Clawson
Johns Hopkins University School of Medicine
lclawson@jhmi.edu

Victoria Chovaz
Johns Hopkins University School of Nursing
vchovaz1@jhmi.edu

Laureano Moro-Velázquez
Johns Hopkins University
laureano@jhu.edu

Abstract

Spoken language understanding (SLU) systems often exhibit suboptimal performance in processing atypical speech, typically caused by neurological conditions and motor impairments. Recent advancements in Text-to-Speech (TTS) synthesis-based augmentation for more fair SLU have struggled to accurately capture the unique vocal characteristics of atypical speakers, largely due to insufficient data. To address this issue, we present a novel data augmentation method for **aty**pical speakers by finetuning a **TTS** model, called Aty-TTS. Aty-TTS models speaker and atypical characteristics via knowledge transferring from a voice conversion model. Then, we use the augmented data to train SLU models adapted to atypical speech. To train these data augmentation models and evaluate the resulting SLU systems, we have collected a new atypical speech dataset containing intent annotation. Both objective and subjective assessments validate that Aty-TTS is capable of generating high-quality atypical speech. Furthermore, it serves as an effective data augmentation strategy, contributing to more fair SLU systems that can better accommodate individuals with atypical speech patterns.

1 Introduction

Atypical speech refers to speech patterns that deviate from typical development or the commonly accepted speaking norms for a particular age, region, or culture. Common speech applications, like automatic speech recognition (ASR) and spoken language understanding (SLU), often exhibit

NeurIPS 2023 Workshop on Synthetic Data for Empowering ML Research.

suboptimal performance when it comes to processing atypical speech (1; 2; 3). To address these performance limitations, two prevalent strategies have gained attention: transfer learning (4) and data augmentation (5; 6). This paper specifically concentrates on the application of data augmentation through Text-to-Speech (TTS) synthesis of atypical speech.

TTS for atypical speakers is a largely unexplored frontier. Soleymanpour *et al.* (7) fine-tuned a TTS model that was initially trained on typical speakers, and added explicit mechanisms for dysarthria severity levels and pause insertion. However, their approach was limited to four coarse-grained dysarthria severity levels, making it less generalizable to other datasets. Matsuzaka *et al.* (6) and Zhao *et al.* (8) generated dysarthric speech by chaining TTS models with Voice Conversion (VC) models, but the speech and pause rates of the target speakers were determined by those of typical speakers, which does not always yield realistic atypical speech. Moreover, previous research efforts have commonly evaluated ASR performance using restricted datasets including UASpeech (9), which only contains isolated words; Torgo (10), which only features 8 dysarthric speakers; or Euphonia (11), which is nonpublic. Notably, none of these studies have assessed the impact of TTS-based data augmentation on SLU performance for atypical speech or conducted subjective evaluations by experts that could indicate the quality of the synthesized speech. While some studies have employed VC techniques to generate atypical speech (12; 13; 14), TTS allows for the modeling of pause, speech rates, and phoneme duration for a particular target speaker or group, without the necessity for source audio (15). This capability enables the convenient creation of synthetic atypical speech tailored to specialized domains, such as complex SLU scenarios with a large number of intents, thereby improving the fairness of SLU systems.

In this paper, we present the following contributions: (1) introduce a novel dataset of atypical speech, HeyJay, specifically curated for SLU. (2) propose a data augmentation method to enhance the modeling of speaker characteristics and atypical characteristics for TTS. (3) evaluate the new method in two SLU scenarios with atypical speech. To our knowledge, this is the first study that evaluates atypical speech with Fluent Speech Commands (FSC) (16) and SLURP (17) intents. (4) conduct subjective evaluations by two expert speech and language pathologists to ascertain whether the synthesized atypical speech retains the distinctive traits of the original speech it aims to emulate.[1]

A diffusion-based TTS model (18) is adapted to atypical speech in this work, although this method can be easily transferred to other TTS models. The idea of the adaptation, called Aty-TTS, is to transfer the knowledge from an already validated VC model for atypical speech to a TTS model. We generate typical-atypical paired data with VC and force the decoder of the TTS model to accomplish auxiliary VC tasks from typical speech to atypical speech. The results on HeyJay show that Aty-TTS mimics characteristics of atypical speech and can significantly improve atypical SLU. This improvement is more pronounced when Aty-TTS is combined with other data augmentation techniques (19; 20).

2 Materials and methods

2.1 New HeyJay corpus and other materials

HeyJay is a new ongoing corpus containing atypical speech. The recorded sentences start with the wakeword "Hey, Jay", which gives the name to the corpus. The goal of HeyJay is to provide the scientific community and software developers with a new corpus of atypical speech, including annotated transcriptions and intent to enable further research and more fair, accessible, and robust spoken-language technologies. We use HermeSpeech (21) to collect atypical speech recordings at hospitals or the participant's homes. The data collection was approved by an Institutional Review Board, and all participants signed an informed consent and were compensated for their collaboration. All the speakers have dysarthria, caused by neurological conditions or cerebrovascular accidents, including Parkinson's Disease, Spinocerebellar Ataxias, Amyotrophic Lateral Sclerosis, and Stroke. Dysarthria is characterized by the misarticulation of phonemes, slow (hypokinetic) or fast (hyperkinetic) speech rate, monotonous intonation, and dysphonia as main signs. In this study, we employ 17 speakers who read sentences with the same intent as the FSC (16) (HeyJay-FSC partition) and 8 speakers reading sentences extracted from the SLURP (17) dataset (HeyJay-SLURP partition), which is notable for its greater number of entities. The recordings from FSC and SLURP were employed

[1] This research was funded by the JHU + Amazon AI2AI Initiative for Interactive AI. The source code and models are available at https://github.com/WangHelin1997/Aty-TTS. The perceptual evaluations of HeyJay speakers and audio demos are shared to allow further studies: https://wanghelin1997.github.io/Aty-TTS-Demo/.

Table 1: Data statistics of SLU datasets. Entities are expressions that refer to objects in SLU.

	FSC (16)	SLURP (17)	HeyJay-FSC	HeyJay-SLURP
Speakers	97	211	17	8
Audio files	30,043	141,530	3,765	1,922
Duration [h]	19	101.5	7.1	4.2
Average length [s]	2.3	2.9	6.8	8.0
Total Entities	334	16,792	334	1,579

in this study to train the SLU models, along with HeyJay and synthesized atypical speech. Data statistics of FSC, SLURP and HeyJay are shown in Table 1. In this study, we also used LJSpeech (22), a dataset comprising roughly 24 hours of recordings from a single speaker, and LibriTTS (23), which offers 586 hours of audio from 2,456 speakers, to pre-train the TTS and VC models, respectively.

2.2 Baseline methods

Grad-TTS: Grad-TTS (18), whose training scheme is shown in Fig. 1a, is used as our baseline model. Given an input text sequence $\boldsymbol{z_x} \in \mathbf{R}^L$ (phoneme sequence in this paper), the model aims at generating a mel-spectrogram $\boldsymbol{z_y} \in \mathbf{R}^{T \times F}$ where T is the number of acoustic frames and F is the number of mel bins. The encoder $f_{encoder}$ maps $\boldsymbol{z_x}$ into a latent feature sequence $\tilde{\boldsymbol{\mu}} \in \mathbf{R}^{L \times F}$, followed by a duration predictor, $f_{duration}$, which predicts phoneme durations $\hat{\boldsymbol{p}} \in \mathbf{R}^L$. Note that $\tilde{\boldsymbol{\mu}}$ contains one latent feature representation per each phoneme contained in the input $\boldsymbol{z_x}$. An aligner $f_{aligner}$ is then used to transform $\tilde{\boldsymbol{\mu}}$ into a latent mel-spectrogram $\boldsymbol{\mu} \in \mathbf{R}^{T \times F}$. For the i-th phoneme in $\boldsymbol{z_x}$, the aligner expands the corresponding latent feature $\tilde{\mu}_i$ for $\lceil \hat{p}_i \rceil$ times. During training, the hard monotonic alignment (24) is used to ensure that the total length of $\boldsymbol{\mu}$ matches T, that is imposed by the desired output spectrogram $\boldsymbol{z_y}$.

$$\tilde{\boldsymbol{\mu}} = f_{encoder}(\boldsymbol{z_x}; \theta_{encoder}) \tag{1}$$

$$\hat{\boldsymbol{p}} = f_{duration}(\tilde{\boldsymbol{\mu}}; \theta_{duration}) \tag{2}$$

$$\boldsymbol{\mu} = f_{aligner}(\tilde{\boldsymbol{\mu}}; \hat{\boldsymbol{p}}) \tag{3}$$

Here, $\theta_{encoder}$ and $\theta_{duration}$ are trainable parameters of the encoder and duration predictor. A decoder $f_{decoder}$ is then used to refine the latent mel-spectrogram and estimate the target mel-spectrogram.

$$\hat{\boldsymbol{z}}_{\boldsymbol{y}} = f_{decoder}(\boldsymbol{\mu}; \theta_{decoder}) \tag{4}$$

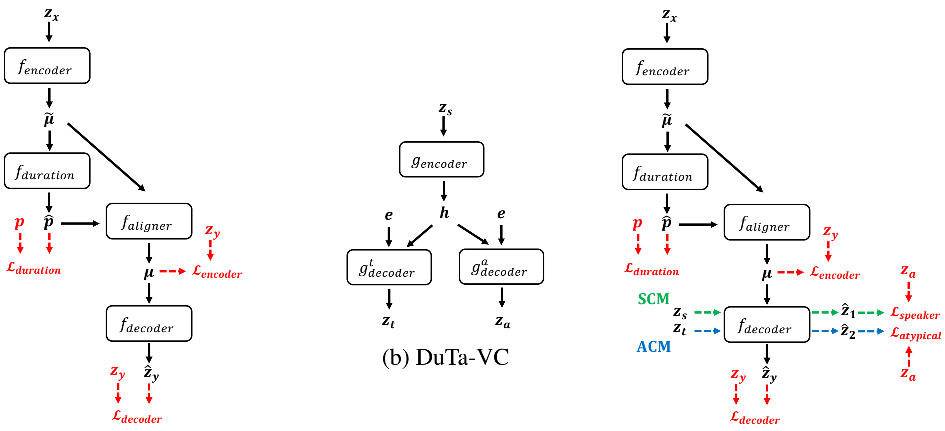
where $\hat{\boldsymbol{z}}_{\boldsymbol{y}} \in \mathbf{R}^{T \times F}$ is the estimated mel-spectrogram and $\theta_{decoder}$ are trainable parameters of the decoder. There are three loss functions in Grad-TTS: encoder loss, duration predictor loss and decoder loss. The encoder loss is applied to minimize the distance between the aligned encoder output $\boldsymbol{\mu}$ and target mel-spectrogram $\boldsymbol{z_y}$. The duration predictor loss is applied to minimize the distance between the predicted duration of the phonemes with the real phoneme duration $\boldsymbol{p} \in \mathbf{R}^L$ in the logarithmic domain. The decoder loss is used to minimize the difference between the decoder output $\hat{\boldsymbol{z}}_{\boldsymbol{y}}$ and target mel-spectrogram $\boldsymbol{z_y}$. The mean square error (MSE) criterion is applied to all of them.

$$\mathcal{L}_{encoder} = \frac{1}{T \times F} \|\boldsymbol{\mu} - \boldsymbol{z_y}\|_2^2 \tag{5}$$

$$\mathcal{L}_{duration} = \frac{1}{L} \|\hat{\boldsymbol{p}} - \boldsymbol{p}\|_2^2 \tag{6}$$

$$\mathcal{L}_{decoder} = \frac{1}{T \times F} \|\hat{\boldsymbol{z}}_{\boldsymbol{y}} - \boldsymbol{z_y}\|_2^2 \tag{7}$$

DuTa-VC: DuTa-VC is a method validated in (12) that provides successful voice conversion from typical to atypical speech. As shown in Fig. 1b, for an input source typical mel-spectrogram $\boldsymbol{z}_s \in \mathbf{R}^{T' \times F}$, the output depends on the decoder, and can be either a typical mel-spectrogram with target speaker timbre $\boldsymbol{z_t} \in \mathbf{R}^{T' \times F}$ via typical-to-typical VC or an atypical mel-spectrogram with the same target speaker timbre $\boldsymbol{z_a} \in \mathbf{R}^{T' \times F}$ via typical-to-atypical VC, where T' denotes the number of acoustic frames. They share the same encoder $g_{encoder}$, where the difference is using a decoder $g^t_{decoder}$ trained with typical speech data or using a decoder $g^a_{decoder}$ trained with atypical speech data. Both decoders have the same input, *i.e.* the hidden representation $\boldsymbol{h}$ from the encoder and a speaker embedding $\boldsymbol{e}$ obtained with a model pre-trained for speaker recognition (25). In our experiments, $g_{encoder}$, $g^t_{decoder}$ and $g^a_{decoder}$ are trained on LibriTTS. $g^a_{decoder}$ is further finetuned on each speaker of HeyJay separately. We do not change phoneme duration like (12) for simplification.



(a) Grad-TTS

(b) DuTa-VC

(c) Aty-TTS

Figure 1: Training scheme of Grad-TTS (a) and Aty-TTS (c), and inference scheme of DuTa-VC (b).

2.3 Aty-TTS

Trying to fine-tune Grad-TTS with atypical speech, speaker and atypical characteristics can be challenging to model, especially for the decoder, as the amount of atypical audio data is often small. In our initial experiments, we found that the synthesized speech mismatched articulation characteristics with the real atypical speech, and the timbre of the synthesized speaker was often not like the target speaker. To overcome these problems, we propose a method to finetune the TTS model, which transfers the knowledge from a VC model to the TTS decoder. It was validated in (26) that voice-converted data more closely matches the target recordings than any other auxiliary data, even when such data originates from a different style or speaker. As a result, the TTS decoder benefits from a broader training set with a consistent distribution.

Consequently, we enhance the training of the TTS decoder by two strategies: speaker characteristics modeling (SCM) and atypical characteristics modeling (ACM), which are shown in Fig. 1c. The idea is to generate typical-atypical paired data with VC and force the TTS model to accomplish auxiliary VC tasks by adapting it with that paired data. More specifically, SCM guides the TTS decoder to convert the source typical mel-spectrogram $\boldsymbol{z_s}$ into the atypical mel-spectrogram with target speaker timbre $\boldsymbol{z_a}$ while ACM guides the TTS decoder to convert the typical mel-spectrogram with target speaker timbre $\boldsymbol{z_t}$ into the atypical mel-spectrogram with target speaker timbre $\boldsymbol{z_a}$. In this way, we can use much more data from VC to train the TTS decoder. While SCM captures both timbre and articulation, we introduced ACM specifically to improve articulation modeling, a crucial element in atypical speech. Our initial experiments indicate that ACM+SCM enhances TTS performance for atypical speakers more than only SCM. The estimated mel-spectrograms $\boldsymbol{\hat{z}_1} \in \mathbf{R}^{T' \times F}$ with SCM and $\boldsymbol{\hat{z}_2} \in \mathbf{R}^{T' \times F}$ with ACM are obtained by

$$\boldsymbol{\hat{z}_1} = f_{decoder}(\boldsymbol{z_s}; \theta_{decoder}) \tag{8}$$

$$\boldsymbol{\hat{z}_2} = f_{decoder}(\boldsymbol{z_t}; \theta_{decoder}) \tag{9}$$

The MSE loss is applied to minimize the difference between the decoder outputs and the atypical mel-spectrogram with target speaker timbre $\boldsymbol{z_a}$.

$$\mathcal{L}_{speaker} = \frac{1}{T' \times F} \|\boldsymbol{\hat{z}_1} - \boldsymbol{z_a}\|_2^2 \tag{10}$$

$$\mathcal{L}_{atypical} = \frac{1}{T' \times F} \|\boldsymbol{\hat{z}_2} - \boldsymbol{z_a}\|_2^2 \tag{11}$$

Aty-TTS is trained with all the loss functions:

$$\mathcal{L} = \mathcal{L}_{encoder} + \mathcal{L}_{duration} + \mathcal{L}_{decoder} + \mathcal{L}_{speaker} + \mathcal{L}_{atypical} \tag{12}$$

We use LibriTTS data as the source audio for DuTa-VC. Aty-TTS models are pre-trained on LJSpeech and then finetuned on HeyJay for each atypical speaker separately. After generating the mel-spectrogram with the decoder, the pre-trained HiFi-GAN vocoder (27) trained on LJSpeech is used to reconstruct the audio waveform. We finetune the vocoder for each atypical speaker separately.

Table 2: Intent Classification Accuracy (ICA) results on HeyJay-FSC and FSC.

Method	HeyJay-FSC			FSC
	Low	High	All	
HuBERT (28)	90.5	77.0	85.4	98.7
+WaveAug (20) + SpecAug (19)	94.0	82.4	90.8	99.0
+WaveAug (20) + SpecAug (19) + Grad-TTS (18)	95.4	85.1	91.9	99.4
+WaveAug (20) + SpecAug (19) + DuTa-VC (12)	**96.1**	89.6	93.6	99.5
+WaveAug (20) + SpecAug (19) + **Aty-TTS (ours)**	95.9	**90.2**	**93.7**	**99.5**

Table 3: SLU F1 results on HeyJay-SLURP and SLURP.

Method	HeyJay-SLURP			SLURP
	Low	High	All	
HuBERT (28)	79.79	68.53	73.69	75.72
+ WaveAug (20) + SpecAug (19)	80.54	72.38	76.13	77.25
+ WaveAug (20) + SpecAug (19) + Grad-TTS (18)	80.71	73.55	77.15	78.60
+ WaveAug (20) + SpecAug (19) + DuTa-VC (12)	79.03	67.50	73.10	73.44
+ WaveAug (20) + SpecAug (19) + **Aty-TTS (ours)**	**80.87**	**75.96**	**78.06**	**79.02**

3 Experiments

In this study, we compare a new data augmentation method that generates synthetic speech, Aty-TTS, with other baseline methods, including WaveAug (20), SpecAug (19), Grad-TTS (18), DuTa-VC (12),. This was done by comparing the performance of different atypical SLU systems trained with these augmented data. Moreover, we further investigate the impact of varying the quantity of synthesized data on SLU training.

3.1 Experiment Setups

Aty-TTS has the same architecture as Grad-TTS (18), and the VC model has the same architecture as DuTa-VC (12). Following their settings, we use 80-dimensional mel-spectrograms with Short-Time Fourier Transform (STFT) window size of 46.4 ms and hop size of 11.6 ms. All the audio files are re-sampled at 22.05 kHz. The number of frames for the Aty-TTS is 172 at training (2 seconds). Adam optimizer is employed with initial learning rates 1×10^{-4} for pre-training and 5×10^{-5} for finetuning, respectively. Batch sizes are set to 64 and 32 with 200 epochs and 50 epochs, respectively. We use 20 hours of VC augmentation for each atypical speaker in Aty-TTS.

3.2 SLU and objective evaluation metrics

SLU with FSC: We used FSC (16) training set and HeyJay-FSC to finetune a HuBERT base model (28), and then test on FSC test set and HeyJay-FSC. Speakers from HeyJay-FSC were categorized into two groups: (i) **Low**-dysarthria (11 speakers) and (ii) **High**-dysarthria (6 speakers)[2]. We divided the 17 speakers randomly into five distinct parts and employed 5-fold cross-validation to conduct leave-speakers-out experiments, where speakers for test were not seen during training. 20 hours of synthesized atypical speech were used, which is similar to the duration of FSC. Following other experiments on FSC (29; 30; 31), intent classification accuracy (ICA) was calculated.

SLU with SLURP: We used SLURP (17) training set and synthesized data with Aty-TTS by all speakers in HeyJay-FSC to finetune a HuBERT base model (28), and then test on SLURP test set and HeyJay-SLURP. We excluded far-range audios from the SLURP test set, as our focus in this work is not on noisy speech. None of the speakers or sentences included in the test set were exposed to the model during the training phase. Test speakers from HeyJay-SLURP were categorized into two groups: (i) **Low**-dysarthria and (ii) **High**-dysarthria, each group containing 4 speakers. We used 100 hours atypical speech synthesized with Aty-TTS to match the duration of SLURP. Following previous studies on SLURP (32; 33; 34), SLU F1 (17) was calculated.

[2] We considered that participants with a dysarthria severity equal or lower than 1.5, in a scale from 0 to 4, had Low-dysarthria severity. Participants with scores higher than 1.5 had High-dysarthria severity.

Table 4: MAE, RMSE and R^2 between the real and synthetic speech on 12 atypical speech traits, including overall dysarthria severity (T1), overall articulation severity (T2), imprecise consonants (T3), prolonged phonemes (T4), irregular breakdowns (T5), distorted vowels (T6), overall voice quality (T7), harsh voice (T8), hoarse/wet voice (T9), breathy voice (T10), strained/strangled voice (T11), stoppages (T12).

Metric	Trait											
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
MAE↓	0.60	0.56	0.57	0.29	0.47	0.44	0.32	0.41	0.34	0.15	0.25	0.01
RMSE↓	0.75	0.73	0.72	0.40	0.58	0.58	0.51	0.48	0.50	0.24	0.32	0.06
R^2↑	0.56	0.57	0.52	0.73	0.48	0.48	0.66	0.08	0.62	0.90	0.82	1.00

3.3 Subjective Evaluation Metrics

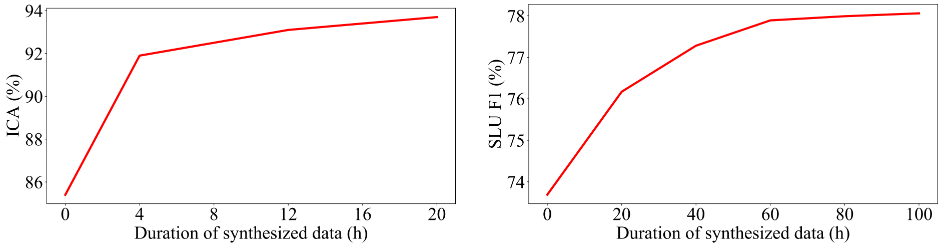
Two speech and language pathologists with more than 10 years of experience blindly evaluated both the real and generated speech samples generated by Aty-TTS. Each pathologist was provided with eight audio clips for each speaker and was asked to assess the overall degree of atypical speech traits. The assessment scores spanned from 0 to 4, where 0 indicated no abnormality and 4 indicated a severe condition, following the Rating Scale for Deviant Speech Characteristics protocol described in chapter three of (35). For each set of real and synthetic speech, we computed the average score for the eight clips from each of the two pathologists. Statistical measures were calculated to compare the differences between the scores of the real and generated speech for each particular trait: mean absolute error (MAE) which provides a mean value representation of the error, root-mean-square error (RMSE) which offers insight into the magnitude of errors and penalizing larger mistakes, and the coefficient of determination (R^2) which reveals the proportion of data variance.

3.4 Results

The SLU performance for the FSC and HeyJay-FSC datasets is detailed in Table 2. To maintain consistency across experiments, we keep factors such as the number of training iterations and the amount of synthesized data constant. Specifically, we finetune a pre-trained Grad-TTS model for each speaker in the HeyJay-FSC subset. Our findings indicate that performing SLU on HeyJay-FSC is considerably more challenging than on the FSC dataset. The impact of high-dysarthria speech on SLU performance is also more pronounced compared to low-dysarthria speech. Signal-based data augmentation techniques (WaveAug and SpecAug) significantly enhance intent classification. Methods relying on synthesized speech further improve performance. Among these, Grad-TTS performs the least effectively. In contrast, both Aty-TTS and DuTa-VC successfully model atypical characteristics, yielding notably better results, especially for speakers with high levels of dysarthria. Table 3 shows the SLU performance for the SLURP and HeyJay-SLURP datasets. Aty-TTS outperforms all other techniques across both dysarthria groups, demonstrating particularly marked improvements for speakers with high levels of dysarthria compared to Grad-TTS. DuTa-VC's performance is suboptimal, due to the noisy and reverberant recordings of the SLURP, which are used as source audios in VC. In such challenging acoustic conditions, TTS-based augmentations prove to be a more effective strategy than VC-based ones. While VC is also employed for augmentation in Aty-TTS training, clean source data can be used for VC, such as LibriTTS. In summary, Aty-TTS is more versatile than VC in these scenarios, and it enhances the fairness of SLU systems by narrowing the performance gap between typical and atypical speech, as well as between speech with low and high levels of dysarthria.

Additionally, we use a range from 0 to 20 h of data synthesized by Aty-TTS for HeyJay-FSC and from 0 to 100 hours for HeyJay-SLURP. As illustrated in Fig.2a, even a modest amount of synthesized data (e.g., 4 h) yields a substantial improvement in performance on the HeyJay-FSC. While additional data does lead to better accuracy, the performance gains plateau, showing marginal improvements beyond 8 h of synthesized data. In contrast, for HeyJay-SLURP, as depicted in Fig.2b, performance continually improves with the inclusion of more synthesized data. This is likely due to the greater complexity of the SLURP dataset compared to FSC and a more fine-gained metric (SLU F1) is applied. An optimal performance is observed when utilizing 60 h of synthesized data.

Table 4 presents the findings from our subjective evaluation metrics. Aty-TTS effectively models various speech characteristics, including articulatory prolonged phonemes, overall voice quality, hoarseness, wetness, breathiness, strained or strangled voice, and stoppages, with MAE below 0.34,



(a) ICA results on HeyJay-FSC

(b) SLU F1 results on HeyJay-SLURP

Figure 2: Influence of different amounts of synthesized data on HeyJay.

RMSE below 0.51, and R^2 over 0.62. These values indicate a strong match in terms of both dysarthria severity and articulation quality between the real and synthetic speech samples.

4 Conclusions

We proposed a data augmentation method to improve TTS for atypical speakers with VC. The TTS model could generate high-quality atypical speech and be an effective data augmentation method for more fair SLU. In future works, we will (1) record more speakers (2) evaluate the method on ASR (3) explore multi-speaker TTS for atypical speakers.

References

[1] Laureano Moro-Velazquez, JaeJin Cho, Shinji Watanabe, Mark A Hasegawa-Johnson, Odette Scharenborg, Heejin Kim, and Najim Dehak, “Study of the performance of automatic speech recognition systems in speakers with parkinson’s disease,” *Proc. INTERSPEECH*, pp. 3875–3879, 2019.

[2] Seyed Reza Shahamiri, “Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 852–861, 2021.

[3] Pu Wang, Bagher BabaAli, and Hugo Van hamme, “A study into pre-training strategies for spoken language understanding on dysarthric speech,” in *Proc. INTERSPEECH*, 2021, pp. 36–40.

[4] Feifei Xiong, Jon Barker, Zhengjun Yue, and Heidi Christensen, “Source domain data selection for improved transfer learning targeting dysarthric speech recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7424–7428.

[5] Mengzhe Geng, Xurong Xie, Shansong Liu, Jianwei Yu, Shoukang Hu, Xunying Liu, and Helen Meng, “Investigation of data augmentation techniques for disordered speech recognition,” in *Proc. INTERSPEECH*, 2020, pp. 696–700.

[6] Yuki Matsuzaka, Ryoichi Takashima, Chiho Sasaki, and Tetsuya Takiguchi, “Data augmentation for dysarthric speech recognition based on text-to-speech synthesis,” in *IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech)*. IEEE, 2022, pp. 399–400.

[7] Mohammad Soleymanpour, Michael T Johnson, Rahim Soleymanpour, and Jeffrey Berry, “Synthesizing dysarthric speech using multi-speaker tts for dysarthric speech recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7382–7386.

[8] Yunxin Zhao, Minguang Song, Yanghao Yue, and Mili Kuruvilla-Dugdale, “Personalizing tts voices for progressive dysarthria,” in *IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2021, pp. 1–4.

[9] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S Huang, Kenneth Watkin, and Simone Frame, “Dysarthric speech database for universal access research,” in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[10] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff, “The torgo database of acoustic and articulatory speech from speakers with dysarthria,” *Language Resources and Evaluation*, vol. 46, pp. 523–541, 2012.

[11] Robert L. MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A. Ladewig, Jimmy Tobin, Michael P. Brenner, Philip C. Nelson, Jordan R. Green, and Katrin Tomanek, “Disordered speech data collection: Lessons learned at 1 million utterances from project euphonia,” in *Proc. INTERSPEECH*, 2021, pp. 4833–4837.

[12] Helin Wang, Thomas Thebaud, Jesús Villalba, Myra Sydnor, Becky Lammers, Najim Dehak, and Laureano Moro-Velazquez, “Duta-vc: A duration-aware typical-to-atypical voice conversion approach with diffusion probabilistic model,” in *Proc. INTERSPEECH*, 2023, pp. 1548–1552.

[13] Disong Wang, Jianwei Yu, Xixin Wu, Songxiang Liu, Lifa Sun, Xunying Liu, and Helen Meng, “End-to-end voice conversion via cross-modal knowledge distillation for dysarthric speech reconstruction,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7744–7748.

[14] Wen-Chin Huang, Bence Mark Halpern, Lester Phillip Violeta, Odette Scharenborg, and Tomoki Toda, “Towards identity preserving normal to dysarthric voice conversion,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6672–6676.

[15] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, “Fastspeech: Fast, robust and controllable text to speech,” *Advances in neural information processing systems*, vol. 32, 2019.

[16] Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio, “Speech model pre-training for end-to-end spoken language understanding,” *Proc. INTERSPEECH*, pp. 814–818, 2019.

[17] Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser, “SLURP: A spoken language understanding resource package,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 7252–7262, Association for Computational Linguistics.

[18] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov, “Grad-tts: A diffusion probabilistic model for text-to-speech,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 8599–8608.

[19] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, “Specaugment: A simple data augmentation method for automatic speech recognition,” *Proc. INTERSPEECH*, pp. 2613–2617, 2019.

[20] Eugene Kharitonov, Morgane Rivière, Gabriel Synnaeve, Lior Wolf, Pierre-Emmanuel Mazaré, Matthijs Douze, and Emmanuel Dupoux, “Data augmenting contrastive learning of speech representations in the time domain,” in *IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 215–222.

[21] Jang-Woo Park, Maximilian Zinkus, Jim Huang, Ankur Butala, Jayne Zhang, Lora Clawson, Sarah Cust, Victoria Chovaz, Helin Wang, and Laureano Moro-Velazquez, “Hermespeech recorder: a new open-source web platform to record speech to the cloud,” in *Proceedings of MAVEBA*, 2023.

[22] Keith Ito and Linda Johnson, “The lj speech dataset,” `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[23] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, “Libritts: A corpus derived from librispeech for text-to-speech,” *Proc. INTERSPEECH*, pp. 1526–1530, 2019.

[24] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon, “Glow-tts: A generative flow for text-to-speech via monotonic alignment search,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.

[25] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al., “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” *Advances in neural information processing systems*, vol. 31, 2018.

[26] Goeric Huybrechts, Thomas Merritt, Giulia Comini, Bartek Perz, Raahil Shah, and Jaime Lorenzo-Trueba, “Low-resource expressive text-to-speech using data augmentation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6593–6597.

[27] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 17022–17033, 2020.

[28] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[29] Loren Lugosch, Brett H Meyer, Derek Nowrouzezahrai, and Mirco Ravanelli, "Using speech synthesis to train end-to-end spoken language understanding models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8499–8503.

[30] Shu Wen Yang, Po Han Chi, Yung Sung Chuang, Cheng I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan Ting Lin, et al., "Superb: Speech processing universal performance benchmark," in *Proc. INTERSPEECH*, 2021, pp. 3161–3165.

[31] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. INTERSPEECH*, 2021, pp. 1194–1198.

[32] Siddhant Arora, Siddharth Dalmia, Pavel Denisov, Xuankai Chang, Yushi Ueda, Yifan Peng, Yuekai Zhang, Sujay Kumar, Karthik Ganesan, Brian Yan, et al., "Espnet-slu: Advancing spoken language understanding through espnet," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7167–7171.

[33] Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe, "Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding," in *International Conference on Machine Learning*. PMLR, 2022, pp. 17627–17643.

[34] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[35] Joseph R Duffy, *Motor Speech disorders-E-Book: Substrates, differential diagnosis, and management*, Elsevier Health Sciences, 2012.