Advancing Pediatric ASR: The Role of Voice Generation in Disordered Speech

Karen Rosero[1], *Ali N. Salman*[2,3], *Shreeram Chandra*[3,4], *Berrak Sisman*[4], *Cortney Van't Slot*[5], *Alex A. Kane*[5], *Rami R. Hallac*[5,6], *Carlos Busso*[1]

[1]Language Technologies Institute, Carnegie Mellon University, PA, USA; [2]ARRAY Innovation, BH; [3]ECE, The University of Texas at Dallas, TX, USA; [4]CLSP, Johns Hopkins University, MD, USA; [5]Department of Plastic Surgery, University of Texas Southwestern Medical Center, TX, USA; [6]Analytical Imaging and Modeling Center, Children's Health, TX, USA;

kroseroj@cmu.edu, sisman@jhu.edu, rami.hallac@childrens.com, busso@cmu.edu

Abstract

Zero-shot performance of state-of-the-art *automatic speech recognition* (ASR) significantly declines on pediatric patients with *speech sound disorders* (SSDs) due to deviations in phonetic pronunciation. To address this, we train a subject-agnostic ASR system on 77 minutes of pediatric SSD transcribed data, which improved zero-shot ASR by 67.48%. Given the scarcity of data and privacy concerns with children's data, we study the suitability of *voice conversion* (VC) and *text-to-speech* (TTS) to synthesize disorder-reflective samples. Our ASR system surpassed zero-shot by 71.72% when leveraging TTS and showed potential for privacy preservation when using VC. Notably, pretraining on synthetic samples alone reduces the required real SSD data to 50 minutes (i.e., 65% of the data), while achieving metrics comparable to the model finetuned with all SSD samples. This study enables ASR technologies to assist individuals with SSDs and facilitates automatic transcription of speech therapy sessions.

Index Terms: speech disorders, automatic speech recognition, voice conversion, children speech.

1. Introduction

Speech is fundamental to human communication, but its effectiveness can be impeded when individuals face challenges in producing or articulating speech sounds [1, 2]. SSDs include motor-speech disorders (e.g., apraxia or dysarthria), structural anomalies (e.g., cleft lip and palate or craniofacial differences), fluency and voice disorders [3]. These conditions prevail in early childhood [4]. Abnormal unintelligibility in pediatric patients with SSDs significantly impacts their social interactions, emotional well-being, and overall quality of life [5].

Learning-based systems for face and speech analysis exhibit limited performance for individuals with SSD [6]. This is particularly true for ASR systems, which underperform because they are trained predominantly on *typically developed* (TD) adult speech, failing to capture the pronunciation variability in children with SSDs [7]. Additionally, diagnosis and treatment of SSDs rely on auditory perceptual analyses conducted by speech-language therapists, which are prone to intra- and inter-rater variability and require significant time and expertise from clinicians [8, 9]. A reliable ASR system tailored for pediatric patients with SSDs could reduce the need for manual transcription and support more standardized approaches to speech evaluation. However, the scarcity of publicly available corpora focused on children with SSDs poses a significant challenge, depraving the system generalizability.

We explore the potential of speech generation techniques to synthesize pediatric disordered speech; specifically, we explore VC and TTS methods. As a baseline, we finetune the speech foundation model Whisper [10] to enhance ASR performance on speech from children with SSDs. We then compare this baseline to models trained on real and synthetic SSD samples. Additionally, we adopt a two-stage progressive fine-tuning approach. The first stage trains the ASR model exclusively on synthetic samples. The second stage further fine-tunes the model using real SSD data, leveraging the synthetic-based model as a starting point. This method aims to reduce the amount of real SSD data required for effective training of an ASR system.

Our approach to synthesizing SSD speech using TTS involves controlled guidance of style acoustic features from SSD recordings, capturing characteristic pronunciation deviations. During synthesis, the generated SSD sample is also influenced, to a lesser extent, by the target lexical information. This ensures the synthesized speech retains the acoustic style of the source SSD utterance while preserving natural prosody. For VC, the lexical content of the generated SSD sample is determined solely by the source utterance. We use real SSD samples as the source and incorporate disentangled speaker identity features from children without SSDs. VC effectively modifies speech characteristics such as timbre and prosody while preserving the disordered pronunciation of the original SSD utterance. We explore ways to leverage TTS and VC techniques to improve SSD-ASR performance.

Our results show that training a subject-independent ASR system with 77 minutes of transcribed SSD audio improves WER by 67.49% over Whisper's zero-shot performance. Using real and TTS-based SSD further enhanced ASR to 70.64%. Progressive finetuning with synthetic TTS followed by real SSD reduced the need for real data to 50 minutes. Training with the VC strategy achieves a WER of 37.34%, similar to the baseline, and improves to 33.74% with progressive finetuning. Since VC still requires real source SSD samples to synthesize the desired linguistic content, it holds potential for privacy protection, but its use case for improved ASR is currently limited. This study highlights the promise of voice generation techniques for augmenting ASR systems for pediatric SSD patients, improving speech assessment while reducing manual effort.

2. Related works

Recent advancements in ASR have largely benefited from large-scale deep learning models. However, these improvements often neglect speakers with atypical speech patterns. Efforts to adapt pre-trained models through fine-tuning on sparse datasets have included adult speakers with speech disorders such as amyotrophic lateral sclerosis [11], dysarthria [12, 13, 14, 15], and impairments caused by cerebral palsy, Parkinson's disease, hearing loss, or ataxia [16, 17]. For pediatric populations, pretrained models have been adapted to assess reading miscues

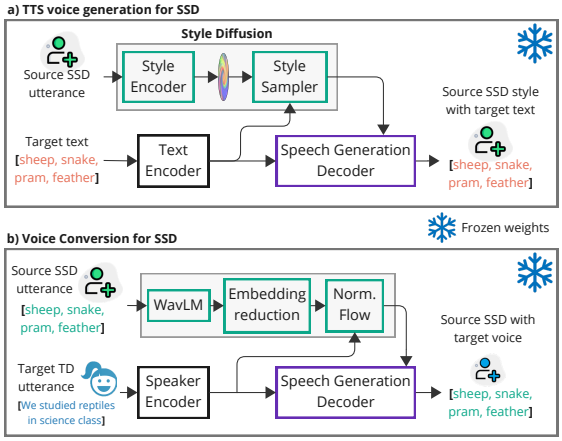

Figure 1: *TTS voice generation and voice conversion pipelines used as data augmentation techniques for improved ASR on pediatric SSD patients.*

[18], mitigate age and speaker bias [19], and evaluate pronunciation [20], with applications to early autism diagnosis [21].

Given the scarcity of datasets representing atypical speech, DA has been explored to enhance ASR performance. Non-generative DA techniques, such as vocal tract length normalization, have been applied to improve children's ASR [22]. In contrast, generative DA methods have been explicitly trained to synthesize speech from typically developing children or adults with dysarthria. Zhang et al. [23] improved child speech recognition through child-to-child voice conversion, while Matsuzaka et al. [24], Hermann et al. [14], and Leung et al. [15] leveraged TTS for dysarthric speech recognition. Additionally, Jin et al. [12] trained an adversarial generative model to synthesize dysarthric speech. In contrast, our work explores the one-shot performance of systems pre-trained on large speech corpora to generate pediatric SSD samples. We analyze the optimal setup and limitations of using pre-trained voice generation techniques, aiming to reduce the need for SSD-specific data.

For children with SSD, DA approaches have mainly been limited to compositional augmentation by mixing words [19]. However, research on voice generation techniques for synthesizing disordered pediatric speech remains scarce. This gap highlights the need for further investigation into robust methodologies for training ASR systems tailored to this population.

3. Methodology

We present our approach to using voice generation techniques to preserve the pronunciation deviation patterns characteristic of pediatric SSD speakers. This section outlines the specific setups implemented for voice conversion and TTS synthesis. Additionally, we detail the preprocessing steps applied to the pediatric SSD corpus and the fine-tuning configurations employed for adapting Whisper to the SSD-ASR task.

3.1. Pediatric SSD corpus

We utilize the UltraSuite corpus [25], which includes ultrasound and acoustic data from child speech therapy sessions. This corpus encompasses three subsets of common childhood speech disorders, such as phonological delay, phonological disorder, inconsistency, vowel and articulation disorders, and childhood apraxia of speech. Participants in the sessions range from 5 to 12 years old. They articulate words, non-word phonemes, and sentences. For our study, we focus exclusively on words and short phrases with available ground-truth transcriptions. We used data from subjects 1 and 4 from the UXSSD subset, subjects 1–20 from UPX, and subjects 1–37 from UX2020. Since the sessions involve speech from both the *speech-language therapist* (SLT) and the child participants, discrepancies between the provided transcriptions and the actual audio content are common. Therefore, we manually corrected the transcriptions and timestamps for 2,429 recordings by removing SLT speech segments and incorporating repeated words into the transcripts where necessary. Praat [26] was used to visualize the frequency representation of utterances and export timestamps and transcriptions. We refer to these recordings as SSD data.

The sentences are designed to evaluate speech production rather than forming coherent messages. The training set contains 77 minutes of speech spoken by 80 subjects consisting of 1,492 utterances, with 429 unique sentences. The development set comprises 30 minutes of speech from three subjects consisting of 552 utterances, with 160 unique sentences. The testing set contains 20 minutes of speech from five subjects, with 388 utterances and 34 unique transcript occurrences. No speakers were shared across our training, development, and testing sets.

3.2. TTS for SSD data augmentation (TTS-SSD)

We utilize StyleTTS2 [27] trained on LibriTTS for TTS generation in a one-shot speaker adaptation setup. The key objective is to include acoustic and lexical patterns similar to those observed in the SSD data. The model requires a source speaker recording of at least three seconds to capture the speaking style. As depicted in Figure 1a, speech style is modeled through prosody, acoustics, and pitch, which are extracted from the source SSD utterance and represented as a latent random variable. This style extraction process captures pronunciation errors by learning the acoustic characteristics of the source speech. During generation, the target lexical content is encoded at the phoneme level using acoustic and prosodic text encoders. These encoders, together with the latent style variable, guide the diffusion sampler to predict prosody and acoustic representations for the target text. Finally, a GAN-based decoder synthesizes the waveform by combining the style and text representations.

The parameters α and β (ranging from 0 to 1) control the degree of guidance from the source style or the target text. We set $\beta = 0.4$ to maintain similarity to the source style while allowing natural prosody based on the target text. We set $\alpha = 0$ to ensure the synthesized speech retains the acoustic characteristics of the source SSD sample. For generation, we use five diffusion steps and an embedding scale of 1. We consider only four source SSD recordings (2 female, 2 male) from the training set, chosen for their low noise levels and representative speech disorder characteristics. For the target texts, we select sentences with a single occurrence in the corpus. We synthesize one audio per sentence by randomly assigning one of the four extracted speaking styles from the SSD samples. By synthesizing only unique text occurrences, each one leveraging the style of one of the four source speakers with SSD, our strategy contributes to a more linguistically balanced augmented dataset.

3.3. Voice Conversion for SSD augmentation (VC-SSD)

We employ FreeVC [28] trained on VCTK, a one-shot voice conversion system that generates speech with the vocal characteristics of a target speaker while preserving the linguistic content of a source utterance, without requiring text guidance. Since the synthesized utterance retains the linguistic content of the source, and the word sequences in the SSD corpus are

designed for speech therapy, we use SSD recordings from the training set as source utterances. As target utterances, we use samples from typically developing children in the MyST dataset [29] (Figure 1b). Through experimentation, we observed that FreeVC effectively modifies speech characteristics such as timbre and prosody while maintaining the disordered speech pronunciation of the SSD source utterance. This behavior enables the use of the generated samples to train an ASR model for SSD.

The model disentangles linguistic content from the SSD source utterance by compressing feature representations extracted using WavLM [30], a self-supervised pre-trained model for speech-related tasks, in conjunction with a normalizing flow block. A TD child speech is set as the target utterance. This sentence is processed using a pre-trained speaker encoder based on a convolutional neural network bi-directional long short-term memory architecture with attention layers, which outputs a speaker embedding. Finally, the speech generation decoder is a conditional variational autoencoder pre-trained for TTS.

3.4. ASR model fine-tuning

By leveraging Whisper's pre-trained on a large corpus of typically developed adult speech, we aim to transfer its knowledge to recognize general speech patterns. After finetuning with SSD data, we expect to improve its recognition of pronunciation deviations observed in our target population. All utterances used for finetuning are resampled to 16KHz and padded to a fixed length of 30 seconds to be aligned with the architecture's processing constraints. The feature extractor transforms raw audio into log-mel spectrograms comprising 80 frequency bins computed in 25-millisecond windows with a 10-millisecond stride. The transcripts undergo text normalization to remove upper case letters and special characters and are processed by a pretrained BPE tokenizer. During finetuning, weights of both the encoder and decoder are retrained since the encoder needs to be adjusted to the different acoustic characteristics of SSD utterances, and the decoder should learn the specific sequences of words used for speech assessment, which are not commonly found in natural speech. Additionally, the list of tokens excluded during pre-training is reset before fine-tuning for better adaptation.

4. Experiments and Results

Our experiments investigate the effectiveness and suitability of speech generation for pediatric SSD. We evaluate multiple strategies by assessing their ability to enhance ASR on pediatric SSD speech. This section details the experimental setups and presents an analysis of the results for each experiment. The experiments include evaluating Whisper's zero-shot performance on SSD data, single-stage fine-tuning with real and/or synthetic SSD data, and a progressive fine-tuning approach that leverages a checkpoint pre-trained on synthetic SSD data.

4.1. Whisper Zero-Shot Performance on SSD (Exp 1)

The first experiment is to evaluate different Whisper model sizes for zero-shot performance on pediatric SSD speech. As shown in Table 1 for Exp 1, all five Whisper models, varying in number of parameters, exhibited extremely poor performance on pediatric SSD samples. This result highlights a severe mismatch between the training data – primarily consisting of adult speakers with typical speech – and the pediatric SSD samples used for evaluation. Among the models trained exclusively on English denoted with '.en', Whisper medium.en (*Wm*) achieved the best performance. Further increasing model size by using Whisper large, trained on multilingual speech, did not lead to improvements. Based on these findings, we use the pre-trained weights of *Wm* to initialize all subsequent experiments.

Table 1: *Performance metrics for ASR on the testing set of the pediatric SSD corpus. The right arrow → denotes finetuning.*

Model	WER	CER
Exp 1: Zero-shot performance		
Whisper tiny.en	140.69	132.72
Whisper base.en	166.72	129.77
Whisper small.en	126.59	90.65
Whisper medium.en *'Wm'*	**112.99**	**77.33**
Whisper large	119.96	81.86
Exp 2: Single finetuning on synthetic samples		
Wm → TTS	87.93 ± 0.81	36.06 ± 1.74
Wm → TTS-SSD	80.06 ± 0.74	46.38 ± 2.41
Wm → VC-SSD	37.34 ± 0.39	20.40 ± 0.30
Exp 3: Single finetuning on real samples		
Wm → SSD *'Baseline'*	36.75 ± 0.41	19.76 ± 1.08
Wm → TD-Child	141.63 ± 1.94	100.81 ± 1.22
Exp 4: Single finetuning on real and synthetic samples		
Wm → (SSD + VC-SSD)	36.29 ± 0.25	18.96 ± 0.13
Wm → (SSD + TTS)*	35.34 ± 0.25	17.78 ± 0.32
Wm → (SSD + TTS-SSD)*	**33.19 ± 0.25**	**16.94 ± 0.07**
Exp 5: Progressive finetuning		
Wm → TD-Child → SSD	40.35 ± 0.10	21.10 ± 0.13
Wm → VC-SSD → SSD*	33.74 ± 0.17	18.26 ± 0.17
Wm → TTS-SSD → SSD*	**31.62 ± 0.34**	**16.35 ± 0.25**

SSD: Sound Speech Disorder, VC: Voice Conversion, TTS: Text-to-Speech
TD: Typically Developed. (*) Denotes a statistically significant difference based on McNemar's test compared to the baseline ($\alpha < 0.05$).

4.2. Single Finetuning on Synthetic Samples (Exp 2)

Our second evaluation is to finetune *Wm* with synthetic speech. We start with employing a standard TTS with an adult voice sourced from a subset of the UltraSuite project [31]. All training transcripts of the SSD corpus were synthesized through TTS, allowing the adaption of the *Wm* model to the linguistic content of the task, without introducing pronunciation deviations. Table 1 presents the results of fine-tuning *Wm* on these TTS-generated samples (*Wm* → TTS) (Exp 2). The model shows an improvement of 25.06 WER points compared to zero-shot performance, demonstrating that adapting Whisper to the training linguistic content of SSD assessment is beneficial.

Both the acoustic differences in disordered speech production and the linguistic content for speech therapy require specialized training to improve the poor zero-shot performance. We hypothesize that performance could be further improved by incorporating both the linguistic and pronunciation characteristics of pediatric SSD speech. As explained in Section 3.2, the style extracted from one of four randomly selected SSD source utterances was used to synthesize all training transcripts of the SSD corpus. The row *Wm* → TTS-SSD in Table 1 shows that incorporating pronunciation deviations from SSD utterances in the TTS synthesis, in addition to the specific linguistic content, further improved WER by 7.87 points. Although the performance remains far from sufficient for practical SSD transcription, the model trained with synthetic TTS-SSD samples provides a better starting point for fine-tuning with real data, as it brings parameters closer to the linguistic and acoustic domain of SSD speech by using only four utterances as a reference.

Voice conversion cannot be used to synthesize SSD samples solely from transcripts, as the transformed audio retains the textual content of the source utterance while adopting the acoustic characteristics of the target utterance. To explore the potential of voice conversion for augmenting SSD data, we converted all training samples of the SSD corpus by randomly selecting one of four child voices with typical speech. The results in row *Wm* → VC-SSD of Table 1 show that finetuning *Wm* with only voice-converted samples yields performance metrics only slightly worse than those of the *'Baseline'* model trained exclusively on real SSD utterances (*Wm* → SSD). Although VC-SSD is not directly comparable with TTS-SSD -which do not require real SSD utterances—these findings indicate that voice conversion primarily alters speaker identity while preserving the linguistic and pronunciation characteristics of children with SSD.

4.3. Single Finetuning on Real Samples (Exp 3)

Our third evaluation determines the benefits of finetuning the models with real SSD. We establish a baseline by fine-tuning *Wm* with all available real SSD training samples. The row *Wm* → SSD in Table 1 shows that this model achieves a mean WER of 36.75 on the SSD test set. Additionally, to demonstrate that the observed ASR improvement over zero-shot performance is not solely due to adaptation to pediatric speech but also to the presence of speech disorders, we finetune the *Wm* model using only TD children's speech from the MyST dataset, following the data splits of Fan et al. [32]. The resulting model achieved a WER of 10.18 and a *character error rate* (CER) of 5.78 on the MyST test set. These results are aligned with the dataset benchmarks [32]. However, when this model was used for inference on pediatric SSD test samples the performance is poor, as shown in row *Wm* → TD-Child of Table 1, confirming that an ASR system trained exclusively on TD children's speech does not generalize well to pediatric SSD utterances.

4.4. Single Finetuning on Real & Synthetic Samples(Exp 4)

The fourth experiment evaluates the effectiveness of voice conversion versus TTS while still incorporating real SSD samples in the fine-tuning process. After experimenting with different amounts of augmented data, we found that generating samples only for sentences with a unique occurrence in the training set (429 utterances) contributed more to dataset balance than augmenting the entire training set. As shown in Exp 4 of Table 1, VC did not yield a statistically significant improvement over the baseline. This is primarily because VC preserves the deviated pronunciation of the source SSD speech, while the characteristics transferred from the target utterance mainly contribute to speaker-specific features such as pitch and tone. Consequently, the VC-generated samples are redundant in terms of disordered pronunciation, as the *Wm* model may not rely on voice characteristics for improved ASR performance.

TTS captures a latent variable representing the style of the source speaker but does not reconstruct the linguistic content. Instead, the target text embeddings guide the diffusion and decoding processes, introducing variability in the atypical speech patterns. Additionally, we used standard TTS to synthesize the same target text and finetuned *Wm* → (SSD+TTS). As shown in Exp 4 of Table 1, standard TTS for data augmentation improved the baseline by 1.41 WER points but did not surpass the performance achieved using the TTS-SSD-augmented samples, which reached a WER of 33.19 –improving the baseline by 3.56 WER points. Notably, this improvement required only four utterances of different speakers with SSD as sources.

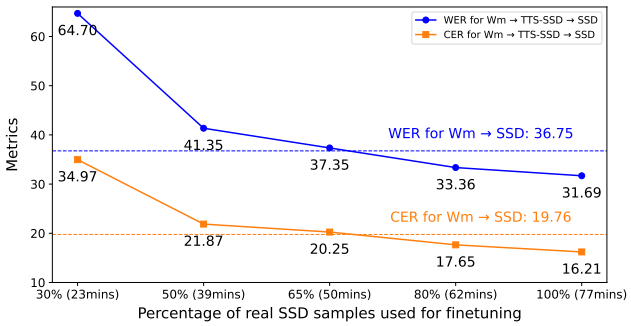

Figure 2: *ASR performance improvement (WER, CER) as a function of the percentage of real SSD data added using the progressive finetuning strategy.*

4.5. Progressive finetuning (Exp 5)

The fifth evaluation further enhances ASR performance on pediatric SSD speech by employing a progressive finetuning strategy. The model is first initialized using a checkpoint trained on synthetic SSD samples, allowing it to adapt to the linguistic content and pronunciation patterns before presenting the real data. Despite achieving a WER of 37.87 in Exp 2, the model trained on synthetic VC-SSD samples shows an additional 4.13-point WER improvement when a second finetuning stage is applied. The best performance is achieved when the model trained on TTS-SSD samples undergoes further finetuning with all real SSD training samples, reaching a WER of 31.62, the best performance observed in this study. This approach, denoted as *Wm* → TTS-SSD → SSD in Table 1, is particularly noteworthy because it required only four real SSD utterances to generate the TTS-SSD set, yet demonstrated the potential of progressive finetuning to surpass the baseline by 5.13 WER points.

Finally, we explore the trade-off between the amount of real SSD data and ASR performance. Figure 2 illustrates WER and CER trends as the amount of real SSD training data is reduced in the progressive finetuning with the *Wm* → TTS-SSD model. A 50% reduction (39 min) increases WER by 4.6 points compared to the baseline. However, retaining 65% of the real SSD data (50 min) achieves a WER of 37.35, only 0.6 points above the baseline, demonstrating that our approach can reduce real data requirements without severely impacting ASR performance.

5. Conclusions

This work explored VC and TTS in generating pediatric speech utterances for ASR training. VC preserves the pronunciation deviations of source SSD speech while altering speaker traits such as pitch and tone. This strategy is promising for speaker anonymization but offers limited benefits for data augmentation. In contrast, TTS captures a source style representation with SSD to guide synthesis, introducing subtle variability in atypical speech patterns as part of the diffusion process, which enhances ASR robustness. Modifying the voice of an SSD speaker while preserving articulation features increases inter-speaker variability in VC, whereas TTS generates novel speech from the same speaker identity, introducing intra-speaker variability. We observe our best ASR performance with a progressive finetuning strategy that starts with a model finetuned with synthetic TTS data followed by real SSD samples, achieving a WER of 31.62. This approach matches the baseline performance using only 50 minutes of SSD data (i.e., 65% of the training data).

6. References

[1] S. A. Borrie, N. Lubold, and H. Pon-Barry, "Disordered speech disrupts conversational entrainment: A study of acoustic-prosodic entrainment and communicative success in populations with communication challenges," *Frontiers in psychology*, vol. 6, p. 1187, 2015.

[2] S.-A. Selouani, M. Sidi Yakoub, and D. O'Shaughnessy, "Alternative speech communication system for persons with severe speech disorders," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 1–12, 2009.

[3] K. Petinou-Loizou, K. Ttofari, and E. Filippou, "What is in a name: Taxonomy of speech sound disorders from a cross-linguistic perspective," *International journal of language & communication disorders*, vol. 59, no. 6, pp. 2123–2130, 2024.

[4] J. Bauman-Waengler and D. Garcia, *Phonological treatment of speech sound disorders in children: A practical guide*. Plural Publishing, 2018.

[5] J. McAllister, J. Skinner, R. Hayhow, J. Heron, and Y. Wren, "The association between atypical speech development and adolescent self-harm," *Journal of speech, language, and hearing research*, vol. 66, no. 5, pp. 1600–1617, 2023.

[6] K. Rosero, A. Salman, B. Sisman, R. Hallac, and C. Busso, "Enhanced facial landmarks detection for patients with repaired cleft lip and palate," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2024)*, Istanbul, Turkey, May 2024, pp. 1–10.

[7] M. Moore, H. Venkateswara, and S. Panchanathan, "Whistle-blowing asrs: Evaluating the need for more inclusive speech recognition systems," *Interspeech 2018*, 2018.

[8] G. P. Usha and J. S. R. Alex, "Speech assessment tool methods for speech impaired children: a systematic literature review on the state-of-the-art in speech impairment analysis," *Multimedia Tools and Applications*, pp. 1–38, 2023.

[9] K. Rosero, A. N. Salman, L. M. Harrison, A. A. Kane, C. Busso, and R. R. Hallac, "Deep learning-based assessment of lip symmetry for patients with repaired cleft lip," *The Cleft Palate Craniofacial Journal*, vol. 0, no. 0, p. 10556656241312730, 2025, pMID: 39838936.

[10] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[11] J. Shor, D. Emanuel, O. Lang, O. Tuval, M. Brenner, J. Cattiau, F. Vieira, M. McNally, T. Charbonneau, M. Nollstadt *et al.*, "Personalizing asr for dysarthric and accented speech with limited data," *Interspeech 2019*, 2019.

[12] Z. Jin, M. Geng, X. Xie, J. Yu, S. Liu, X. Liu, and H. Meng, "Adversarial data augmentation for disordered speech recognition," *arXiv preprint arXiv:2108.00899*, 2021.

[13] J. R. Green, R. L. MacDonald, P.-P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, M. A. Ladewig, J. Tobin, M. P. Brenner *et al.*, "Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases." in *Interspeech*, 2021, pp. 4778–4782.

[14] E. Hermann and M. M. Doss, "Few-shot dysarthric speech recognition with text-to-speech data augmentation," in *Proc. INTERSPEECH 2023*, 2023, pp. 156–160.

[15] W.-Z. Leung, M. Cross, A. Ragni, and S. Goetze, "Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis," *arXiv preprint arXiv:2406.08568*, 2024.

[16] J. Tobin and K. Tomanek, "Personalized automatic speech recognition trained on small disordered speech datasets," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6637–6641.

[17] S. Venugopalan, J. Tobin, S. J. Yang, K. Seaver, R. J. Cave, P.-P. Jiang, N. Zeghidour, R. Heywood, J. Green, and M. P. Brenner, "Speech intelligibility classifiers from 550k disordered speech samples," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[18] L. Gao, C. Tejedor-Garcia, H. Strik, and C. Cucchiarini, "Reading miscue detection in primary school through automatic speech recognition," in *Interspeech 2024*, 2024, pp. 5153–5157.

[19] G. Kim, Y. Eom, S. S. Sung, S. Ha, T.-J. Yoon, and J. So, "Automatic children speech sound disorder detection with age and speaker bias mitigation," in *Interspeech 2024*, 2024, pp. 1420–1424.

[20] Y.-S. Lin, S.-C. Tseng, and J.-S. R. Jang, "Leveraging phonemic transcription and whisper toward clinically significant indices for automatic child speech assessment," in *Interspeech 2024*, 2024, pp. 2440–2444.

[21] J. Li, M. Hasegawa-Johnson, and K. Karahalios, "Enhancing child vocalization classification with phonetically-tuned embeddings for assisting autism diagnosis," in *Interspeech 2024*, 2024, pp. 5163–5167.

[22] Z. Shuyang, M. Singh, A. Woubie, and R. Karhila, "Data augmentation for children asr and child-adult speaker classification using voice conversion methods," in *Proc. Interspeech 2023*, 2023, pp. 4593–4597.

[23] Y. Zhang, Z. Yue, T. Patel, and O. Scharenborg, "Improving child speech recognition with augmented child-like speech," in *Proc. Interspeech 2024*, 2024, pp. 5183–5187.

[24] Y. Matsuzaka, R. Takashima, C. Sasaki, and T. Takiguchi, "Data augmentation for dysarthric speech recognition based on text-to-speech synthesis," in *2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech)*. IEEE, 2022, pp. 399–400.

[25] A. Eshky, M. S. Ribeiro, J. Cleland, K. Richmond, Z. Roxburgh, J. Scobbie, and A. Wrench, "Ultrasuite: A repository of ultrasound and acoustic data from child speech therapy sessions," in *Interspeech 2018*. ISCA, 2018, pp. 1888–1892.

[26] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[27] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, "Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[28] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[29] W. Ward, R. Cole, D. Bolanos, C. Buchenroth-Martin, E. Svirsky, S. V. Vuuren, T. Weston, J. Zheng, and L. Becker, "My science tutor: A conversational multimedia virtual tutor for elementary school science," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 7, no. 4, pp. 1–29, 2011.

[30] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[31] M. S. Ribeiro, J. Sanger, J.-X. Zhang, A. Eshky, A. Wrench, K. Richmond, and S. Renals, "Tal: a synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 1109–1116.

[32] R. Fan, N. Balaji Shankar, and A. Alwan, "Benchmarking children's asr with supervised and self-supervised speech foundation models," in *Interspeech 2024*, 2024, pp. 5173–5177.