End-to-End Dysarthric Voice Conversion for Low-Resource Languages Using Contrastive Soft Units

Anonymous submission to Interspeech 2025

Abstract

Dysarthric voice conversion often relies on cascaded ASR-TTS or text supervision, limiting usability in low-resource languages. Recent unit-based methods remove text dependence but degrade phonetic detail, compromising intelligibility. We propose a textless End-to-End Dysarthric Voice Conversion method that uses soft units-continuous probabilistic representations preserving articulatory nuances. We use a seq2seq Transformer to transform dysarthric speech into more intelligible utterances, trained with cross-entropy and contrastive regularization to enhance robustness across dysarthric severities. A modified HiFi-GAN vocoder synthesizes speech from these units, mitigating mispronunciations from discrete quantization. ASR word error rate improved from 9.71% to 6.16% (mild) and 19.54% to 13.27% (moderate) when trained on resynthesized speech. To our knowledge, this is the first unit-based Tamil dysarthric voice conversion system, achieving intelligibility gains without multitask learning.

Index Terms: Dysarthria, speech resynthesis, End to End system, speech-to-unit translation (S2UT), intelligibility enhancement, automatic speech recognition (ASR), data augmentation

1. Introduction

Dysarthria is a motor speech disorder caused by neurological impairments that affect the muscles responsible for speech production, leading to slowed, slurred, or uncoordinated speech [1]. The severity of dysarthria varies from mild articulation imprecision to severely unintelligible speech, making communication challenging. Automatic Speech Recognition (ASR) systems trained on healthy speech fails to adapt to dysarthric speech due to its acoustic variability and phonetic distortions [2]. This problem is particularly severe in low-resource languages such as Tamil, where publicly available dysarthric speech datasets are scarce, limiting the development of robust ASR and speech resynthesis models.

In recent years, various voice conversion methods have been explored to improve dysarthric speech intelligibility. Earlier approaches include articulatory-based speech modification [3] and statistical parametric speech synthesis [4], which focuses on producing more natural-sounding speech through speaker-adaptive models. Traditional voice conversion pipelines employed Gaussian Mixture Models or Hidden Markov Models to transform dysarthric speech towards a healthier acoustic space. Previous work has introduced spectral mapping strategies [5] and statistical models [6] that improve intelligibility by adapting the acoustic features of dysarthric speech to those of a healthy speaker. More recent deep learning methods have transitioned to neural architectures. Deep neural networks [7] refined phoneme pronunciation via non-linear mappings. Eigenvoice-based VC [8] enabled speaker adaptation with minimal data, making it viable for dysarthric speech enhancement.

Although these methods have shown promise, most rely on text supervision or focus on speaker-to-speaker adaptation. In parallel, textless speech-to-speech translation [9, 10] has emerged as a powerful alternative, enabling direct speech transformations without textual intermediaries. Discrete unit-based models [11] have demonstrated that unlabeled audio can be mapped into phoneme-like clusters, suppressing speaker traits while retaining linguistic content. However, such discrete units can introduce quantization errors and degrade phonetic detail, reducing speech intelligibility. Soft units, which represent speech frames via continuous probabilistic embeddings, have recently been proposed to overcome these limitations [12]. Inspired by speech-to-speech translation systems typically used for cross-lingual tasks, we extend their framework to a monolingual, domain-adaptive setting, where dysarthric speech is modeled as the source language and healthy speech as the target language. We propose a textless dysarthric voice conversion framework that maps dysarthric speech to soft units using a seq2seq Transformer, followed by a HiFi-GAN vocoder to synthesize intelligible speech.

The key contributions of this work include:

- Unlike prior discrete-unit voice conversion approaches, we employ continuous probabilistic representations to retain articulatory detail critical for dysarthric speech clarity, thereby avoiding the phonetic degradation caused by quantization.
- A contrastive regularization strategy enhances robustness across dysarthria severity levels without requiring multi-stage or joint training.
- By preserving linguistic content while enhancing intelligibility, our resynthesized speech serves as domain-augmented training data for dysarthric ASR systems, achieving a 37% relative WER reduction .
- To the best of our knowledge, this is the first Tamil textless dysarthric voice conversion system to adopt soft units, demonstrating scalability for languages lacking orthographic resources or large-scale dysarthria corpora.

2. Related Work

Voice conversion (VC) and speech resynthesis techniques have been widely explored to improve dysarthric speech intelligibility. Early statistical parametric synthesis methods often introduced artifacts, reducing the naturalness of speech [4]. More recent approaches leverage neural vocoders for enhanced speech quality. To enhance automatic speech recognition (ASR) for

dysarthric speakers, data augmentation strategies have been extensively studied. Jiao et al. [13] employed GAN-based voice conversion to generate dysarthric-like speech from healthy speakers, increasing ASR training diversity. Other methods, such as spectral perturbations and adaptive TTS-based synthesis, have been used to improve ASR robustness, albeit with the risk of introducing distortions [14]. Building upon this, Celin et al. [15] explored transfer learning-based continuous dysarthric speech recognition, incorporating data augmentation techniques to mitigate data scarcity and enhance ASR performance.

Self-supervised learning (SSL) and unit-based speech representations have emerged as promising alternatives for dysarthric speech processing. Polyak et al. [16] introduced speech resynthesis from discrete disentangled representations, enabling controlled resynthesis by separating content, prosody, and speaker attributes. Similarly, Wang et al. [17] proposed Unit-DSR, a dysarthric speech reconstruction system that applies HuBERT-based discrete unit learning, followed by a unit-based HiFi-GAN vocoder, to normalize dysarthric speech into a healthy reference speaker's voice. Despite its effectiveness, Unit-DSR relies on discrete units, which can introduce quantization errors that degrade phonetic detail. Moreover, it employs a multi-stage fine-tuning strategy, requiring additional training complexity. In contrast, our work eliminates discrete unit dependence by utilizing soft units, which are continuous probabilistic representations that better preserve articulatory nuances. Additionally, we extend these advancements to Tamil, a low-resource language lacking large-scale dysarthric speech corpora, demonstrating the adaptability of our method beyond high-resource settings without requiring text transcriptions.

3. Proposed Methodology

Following a modified approach from [10], we use mHuBERT to discretize target speech and develop a sequence-to-sequence Speech-to-Unit Translation (S2UT) model. Our pipeline comprises four stages, designed to improve dysarthric speech intelligibility through textless, speaker-agnostic soft-unit modeling. Below, we detail each component.

3.1. mHuBERT-147 finetuning and discrete unit extraction

Figure 1 shows the proposed End-to-End Dysarthric Speech Reconstruction Pipeline where the target clean speech is initially fed into a pretrained multilingual HuBERT (mHuBERT-147) model [18], chosen for its ability to learn cross-lingual speech representations via self-supervised masked prediction [19]. To adapt mHuBERT to Tamil, we fine-tune it on a corpus of healthy Tamil speech. After finetuning, k-means clustering with k-means++ initialization is applied to the resulting features to produce a dictionary of discrete units of size $K = 1000$. Each healthy Tamil utterance is quantized frame-by-frame to an integer index d_t:

$$d_t \in \{0, \ldots, K-1\} \tag{1}$$

These discrete units serve as speaker-independent content representations, disentangling phonetic and speaker attributes during clustering [20].

3.2. Soft content encoder

To mitigate the loss of fine phonetic detail from purely discrete labels, we adopt a soft-units approach [12]. Specifically, a small linear projection layer is trained on top of mHuBERT such that, for each frame, it outputs a probability vector s_t:

$$\mathbf{s}_t \in \mathbb{R}^K \tag{2}$$

The objective is to match each soft distribution $\mathbf{s}_t$ to the corresponding one-hot k-means index d_t using cross-entropy. This training encourages the encoder to remove speaker-specific attributes while retaining nuanced articulatory cues crucial for dysarthric speech intelligibility.

3.3. Direct dysarthric speech-to-soft unit translation

We propose a direct speech-to-soft-unit (S2UT) model that accepts dysarthric audio and predicts the corresponding healthy soft units unlike [9, 17], that relies on auxiliary text-based finetuning or discrete units. Let Mel(Dysarthric) denote log-Mel filterbanks from the dysarthric waveform. We train a Transformer encoder-decoder such that:

$$\{\mathbf{s}_t\}_{t=1}^{T} \longleftarrow \text{S2UT}\left(\text{Mel(Dysarthric)}\right) \tag{3}$$

where each $\mathbf{s}_t$ is the healthy soft-unit distribution at frame t. To enhance robustness against dysarthria severity, we introduce contrastive regularization during training the S2UT Model:

For each input utterance, a negative sample $\{\mathbf{s}_t^-\}$ is drawn from another dysarthric speaker's unrelated utterance. The model minimizes:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{contrast}}, \tag{4}$$

where $\mathcal{L}_{\text{contrast}}$ is defined as:

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(\text{sim}(\mathbf{s}_t, \mathbf{s}_t^+))}{\exp(\text{sim}(\mathbf{s}_t, \mathbf{s}_t^+)) + \exp(\text{sim}(\mathbf{s}_t, \mathbf{s}_t^-))} \tag{5}$$

with $\text{sim}(\cdot)$ denoting cosine similarity. SpecAugment [21] with frequency & time masking is applied to encoder inputs for regularization.

3.4. Soft-unit HiFi-GAN vocoder

As shown in Figure 1 we adapt HiFi-GAN [22], a high-fidelity vocoder renowned for its efficiency in waveform generation to synthesize waveforms from soft units. Each soft-unit distribution $\mathbf{s}_t$ is projected to a continuous embedding $\mathbf{z}_t \in \mathbb{R}^D$ via a learnable matrix $\mathbf{E} \in \mathbb{R}^{K \times D}$:

$$\mathbf{z}_t = \mathbf{s}_t \mathbf{E}. \tag{6}$$

The sequence $\mathbf{z}_t$ is upsampled through transposed convolutions and refined via residual blocks. Following [23], we optimize adversarial, feature-matching, and Mel-spectrogram ℓ_1 losses. The key difference from discrete-unit HiFi-GAN is that we no longer rely on a single integer cluster index per frame; each $\mathbf{s}_t$ is a distribution over multiple clusters, preserving subtle articulatory cues and eliminating quantization artifacts.

3.5. Inference pipeline

During inference, dysarthric audio undergoes three stages. First, the S2UT model translates dysarthric Mel spectrograms into healthy soft-unit distributions $\mathbf{s}_t$. Next, these distributions are projected to continuous embeddings $\mathbf{z}_t = \mathbf{s}_t \mathbf{E}$. Finally, the HiFi-GAN vocoder reconstructs the final speech waveform from $\mathbf{z}_t$.

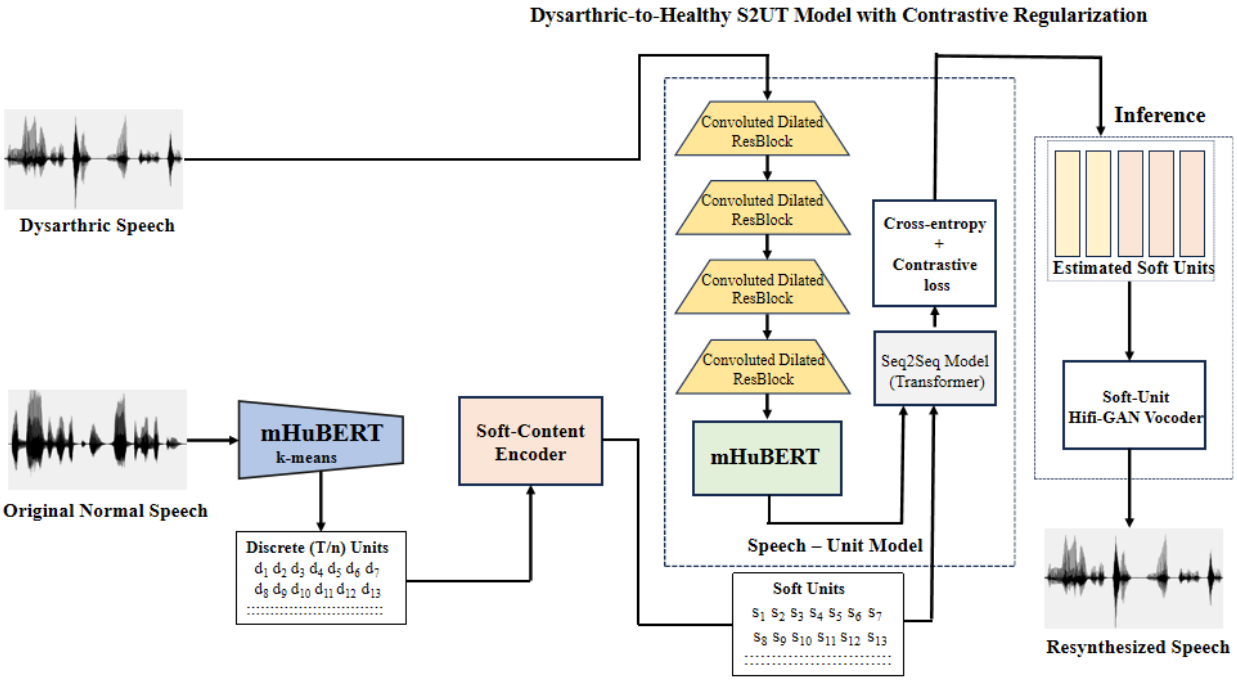

Figure 1: *End-to-End Dysarthric Voice Conversion Pipeline*

4. Experimental setup

4.1. Data

All experiments were conducted using the Tamil Dysarthric Speech Corpus (TDSC) [24, 25], which contains speech recordings from individuals with dysarthria and healthy speakers. The dataset includes 20 dysarthric speakers (13 male, 7 female) diagnosed with cerebral palsy (spastic quadriplegia or bilateral paraplegia) and 10 healthy speakers (5 male, 5 female). Each speaker recorded 365 utterances, comprising 103 isolated words and 262 sentences (ranging from 2 to 6 words), ensuring coverage of all Tamil phonemes. The recordings were collected in collaboration with the National Institute for the Empowerment of Persons with Multiple Disabilities (NIEPMD), using a head-mounted microphone in a controlled laboratory environment at a sampling rate of 16 kHz.

For this study, we used only the mild and moderate dysarthric speakers, totaling 17 dysarthric speakers, while excluding severe cases to facilitate more stable unit alignments. The healthy speaker subset was used to fine-tune our mHuBERT model, providing 3,650 utterances (10 speakers $\times$ 365 utterances). Dysarthric speech served as input to the Speech-to-Unit Translation (S2UT) model, with target outputs derived from the corresponding healthy speaker's soft units.

4.2. System setup

The mHuBERT model used is the pretrained multilingual HuBERT-147 (95M parameters, 12 transformer layers, trained on 147 languages). It includes over 100 hours of Tamil, making it well-suited for adaptation. We fine-tuned mHuBERT on healthy Tamil speech for 50 epochs using AdamW (lr = 3e-5, batch = 32). Discrete units were obtained via k-means++ clustering (K = 1000) on the 11th-layer features, serving as the reference phonetic space for soft-unit modeling. For soft-unit encoding, a linear projection layer was trained on top of the mHuBERT for 25 epochs (Adam, lr = 1e-4, batch = 64) with a cross-entropy loss against discrete unit labels. This preserves speaker invariance while capturing fine phonetic detail.

Our S2UT model follows a Transformer encoder-decoder architecture (12 encoder, 6 decoder layers). The encoder has 4 attention heads, 256-dimensional embeddings, and a Conv1D subsampling module, while the decoder uses 8 attention heads with shared input-output embeddings. Training was conducted for 235 epochs ($\approx$36k steps) using Adam (lr = 5e-4, β_1 = 0.9, β_2 = 0.98, 10k warmup) with cross-entropy and contrastive loss (λ = 0.1) to enhance robustness across dysarthria severities. SpecAugment (frequency masking F = 10, time masking T = 50) was applied for additional regularization. Training took approximately 72 hours. For speech resynthesis, HiFi-GAN was adapted to accept soft-unit inputs. Soft-unit probability vectors were projected into a continuous embedding space ($E \in \mathbb{R}^{1000\times256}$) and processed via transposed convolutions and residual blocks. The vocoder was trained for 5 days (500k steps, batch = 16, Adam, lr = 2e-4) optimizing adversarial, feature-matching, and Mel-spectrogram ℓ_1 losses. All models were implemented in Fairseq and trained on an NVIDIA GeForce RTX 3050 GPU (24 GB VRAM).

4.3. Evaluation experiments

For our experiments, we selected specific subsets from the TDSC dataset, focusing on the Mild (1.42 hrs) and Moderate (2.1 hrs) categories. To build the ASR system, we first constructed monogram and trigram language models using Kaldi, which were then utilized to train a DNN-HMM-based ASR sys-

tem for evaluation. The training set consisted of 292 utterances, while 73 utterances were reserved for testing. To assess the effectiveness of the proposed methodology, we conducted three sets of experiments, comparing the Word Error Rate (WER) of the ASR system across different setups:

- **Baseline ASR System** – The ASR model was trained using only the original dysarthric speech data.
- **Resynthesised ASR System** – The ASR model was trained exclusively on the resynthesized speech generated from our Voice conversion pipeline.
- **Baseline + Resynthesised ASR System** – The ASR model was trained on a combined dataset consisting of both the original dysarthric speech and the resynthesized speech

5. Results

To evaluate the effectiveness of our proposed dysrathric voice conversion method, we conducted subjective and objective assessments using both intelligibility and ASR-based metrics. The evaluation focused on comparing the intelligibility improvements in the resynthesized speech against the original dysarthric speech for both mild and moderate dysarthria cases.

5.1. Intelligibility Assessment

A Mean Opinion Score (MOS) test was conducted with 10 participants who rated the intelligibility of both original and resynthesized speech samples on a 5-point scale. Additionally, Degraded MOS (DMOS) was collected to measure the perceived distortion introduced in the resynthesized speech. Higher MOS scores indicate greater intelligibility, while lower DMOS scores suggest reduced perceived distortion. As shown in Table 1, resynthesized speech achieved higher MOS scores than the original dysarthric speech for both mild and moderate speakers, indicating improved clarity. The MOS for mild dysarthric patients increased from **3.8 to 4.2**, while for moderate dysarthric patients, it improved from **2.8 to 3.5**. However, DMOS results reveal that resynthesized speech exhibited reduced degradation for mild speakers from **3.2** to **2.1**, but for moderate dysarthria, the resynthesized speech had a slightly lower DMOS than the original from **3.6** to **3.3**. This suggests that the proposed methodology successfully enhances the intelligibility of dysarthric speech, making it more comprehensible, particularly for mild and moderate dysarthria. The improved MOS and reduced DMOS scores indicate that the resynthesized speech not only retains key articulatory nuances but also minimizes distortions.

Table 1: *Comparison of Intelligibility Metrics for Original and Resynthesized Speech*

Metric	**Mild**		**Moderate**	
	Original	**Resynthesized**	**Original**	**Resynthesized**
MOS	3.8 ± 0.4	**4.2 ± 0.3**	2.8 ± 0.5	**3.5 ± 0.4**
DMOS	3.2 ± 0.6	**2.1 ± 0.5**	3.6 ± 0.2	**3.3 ± 0.5**
BERT	0.76 ± 0.03	**0.89 ± 0.02**	0.55 ± 0.04	**0.75 ± 0.03**

5.2. Objective Evaluation

To quantitatively assess intelligibility improvements, we computed the BERT-based similarity score which ranges from 0 to 1, measuring how closely the resynthesized speech aligns with the healthy reference speech. As expected, the resynthesized speech exhibited higher BERT scores than the original dysarthric speech, with mild dysarthria improving from **0.76 to 0.89** and moderate dysarthria increasing from **0.55 to 0.75**. These results indicate that the soft-unit transformation effectively enhances phonetic clarity.

Table 2: *Comparison of ASR Performance (WER) for Different Training Setups*

Experimental Setup	**Mild (WER %)**	**Moderate (WER %)**
Baseline	9.71	19.54
Resynthesized	**6.16**	**13.27**
Baseline + Resynthesized	**5.73**	**9.44**

5.3. ASR Performance Results

Table 2 presents the Word Error Rate (WER) comparisons for different training strategies on the Mild and Moderate dysarthric speech categories. The Baseline system, trained on original dysarthric speech, yielded a WER of **9.71%** for mild dysarthria and **19.54%** for moderate dysarthria. However, the Resynthesized system trained on resynthesized speech demonstrated improvements in both categories, reducing WER to **6.16%** and **13.27%**, respectively. The Baseline + Resynthesized system, which combines both original and resynthesized speech, achieved a WER of **5.73%** for mild dysarthria and **9.44%** for moderate dysarthria, indicating that despite the lack of transcription-based supervision, the model is still able to produce competitive ASR performance.

These results show that resynthesized speech enhances intelligibility and can be used for real-time inference to improve communication. Additionally, it serves as valuable augmentation data for ASR, particularly for Mild and Moderate dysarthria, where it significantly reduces WER.

6. Conclusion

This study presents a novel textless dysarthric voice conversion framework that significantly enhances speech intelligibility and ASR performance without relying on transcription. Using self-supervised soft unit modeling and a Transformer-based S2UT pipeline, our approach effectively resynthesizes dysarthric speech into a more intelligible form. MOS, DMOS, and BERT scores confirm perceptual and phonetic clarity improvements, especially for mild and moderate dysarthria. Additionally, the ASR experiments demonstrate that incorporating resynthesized speech as augmentation data reduces WER, showcasing its potential to aid dysarthric speech recognition. Despite these results, challenges remain for the severe dysarthric category, where additional adaptation strategies are required.

Future work will explore speaker-conditioned unit modeling to improve generalization across severity levels. We will also investigate domain adaptation techniques and adversarial learning to enhance soft-unit representations. By expanding this approach to other low-resource languages, we aim to advance research toward more inclusive and accessible dysarthric speech technologies.

7. References

[1] J. R. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. Elsevier Health Sciences, 2019.

[2] M. Tu, M. Nagaraj, and P. N. Garner, "Automatic speech recognition for disordered speech: A review," *Speech Communication*, vol. 78, pp. 1–16, 2016.

[3] F. Rudzicz, T. Hasegawa-Johnson, and J. R. Green, "The torgo database of articulatory and acoustic speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[4] C. Veaux, J. Yamagishi, and S. King, "Speech synthesis for individuals with vocal disabilities: Voice banking and reconstruction," *Speech Communication*, vol. 73, pp. 49–64, 2016.

[5] A. Kain, A. Amano-Kusumoto, and J. P. Hosom, "Spectral voice conversion for enhancing the intelligibility of dysarthric speech," in *Proceedings of ICASSP*, 2007, pp. 988–991.

[6] A. B. T. Toda and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, 2012, pp. 2222–2235.

[7] X. W. L. Sun and X. Xu, "Deep neural network-based voice conversion for dysarthric speech enhancement," in *Proceedings of Interspeech*, 2016, pp. 1542–1546.

[8] T. V. E. Helander and J. Nurminen, "Voice conversion using eigenvoices," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[9] A. Lee, P.-J. Chen, C. Wang, J. Gu, S. Popuri, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang, J. Pino, and W.-N. Hsu, "Direct speech-to-speech translation with discrete units," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3327–3339. [Online]. Available: https://aclanthology.org/2022.acl-long.235/

[10] A. Lee, H. Gong, P.-A. Duquenne, H. Schwenk, P.-J. Chen, C. Wang, S. Popuri, Y. Adi, J. Pino, J. Gu, and W.-N. Hsu, "Textless speech-to-speech translation on real data," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 860–872. [Online]. Available: https://aclanthology.org/2022.naacl-main.63/

[11] Y. A. A. Polyak and N. A. Smith, "Disentangling content and speaker information in speech representations," in *Proceedings of ACL*, 2021, pp. 3456–3468.

[12] B. van Niekerk, M.-A. Carbonneau, J. Zaïdi, M. Baas, H. Seuté, and H. Kamper, "A comparison of discrete and soft speech units for improved voice conversion," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6562–6566.

[13] S. Z. Y. Jiao and J. Tao, "An investigation of gan-based data augmentation for dysarthric speech recognition," in *Proceedings of ICASSP*, 2018, pp. 5119–5123.

[14] L. Vachhani, S. Gupta, and A. Alwan, "Data augmentation using spectral and prosodic perturbations for dysarthric speech recognition," in *Proceedings of Interspeech*, 2018, pp. 2948–2952.

[15] T. A. M. Celin, P. Vijayalakshmi, and T. Nagarajan, "Data augmentation techniques for transfer learning-based continuous dysarthric speech recognition," *Circuits, Systems, and Signal Processing*, vol. 42, no. 1, pp. 601–622, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s00034-022-02156-7

[16] A. Polyak, Y. Adi, and N. A. Smith, "Disentangling content and speaker information in speech representations," in *Proceedings of ACL*, 2021, pp. 3456–3468.

[17] J. Wang, H. Zhang, and X. Liu, "Unit-dsr: Dysarthric speech reconstruction using self-supervised speech units," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 233–247, 2024.

[18] M. Z. Boito, V. Iyer, N. Lagos, L. Besacier, and I. Calapodescu, "mhubert-147: A compact multilingual hubert model," 2024. [Online]. Available: https://arxiv.org/abs/2406.06371

[19] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, and Lang. Proc.*, vol. 29, pp. 3451–3460, October 2021.

[20] W.-C. Huang, Y.-C. Wu, and T. Hayashi, "Any-to-one sequence-to-sequence voice conversion using self-supervised discrete speech representations," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5944–5948.

[21] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019*. ISCA, September 2019. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2680

[22] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[23] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. rahman Mohamed, and E. Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," in *Interspeech*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:262491522

[24] M. Celin, T. Nagarajan, and P. Vijayalakshmi, "Dysarthric speech corpus in tamil for rehabilitation research," 11 2016, pp. 2610–2613.

[25] P. Vijayalakshmi, T. A. Mariya Celin, and T. Nagarajan, "The ssnce database of tamil dysarthric speech," Web Download, Philadelphia, 2021, lDC2021S04. Available at: https://doi.org/10.35111/hkh2-vh40.