Unsupervised Rhythm and Voice Conversion to Improve ASR on Dysarthric Speech

Karl El Hajal[1,2], *Enno Hermann*[1], *Sevada Hovsepyan*[1], *Mathew Magimai.-Doss*[1]

[1]Idiap Research Institute, CH-1920 Martigny, Switzerland
[2]EPFL, École polytechnique fédérale de Lausanne, CH-1015 Lausanne, Switzerland
{karl.elhajal,enno.hermann,sevada.hovsepyan,mathew}@idiap.ch

Abstract

Automatic speech recognition (ASR) systems struggle with dysarthric speech due to high inter-speaker variability and slow speaking rates. To address this, we explore dysarthric-to-healthy speech conversion for improved ASR performance. Our approach extends the Rhythm and Voice (RnV) conversion framework by introducing a syllable-based rhythm modeling method suited for dysarthric speech. We assess its impact on ASR by training LF-MMI models and fine-tuning Whisper on converted speech. Experiments on the Torgo corpus reveal that LF-MMI achieves significant word error rate reductions, especially for more severe cases of dysarthria, while fine-tuning Whisper on converted data has minimal effect on its performance. These results highlight the potential of unsupervised rhythm and voice conversion for dysarthric ASR. Code available at: https://github.com/idiap/RnV.

Index Terms: Rhythm Modeling, Voice Conversion, Unsupervised, Dysarthric Speech Recognition

1. Introduction

Motor speech impairments like dysarthria can significantly hinder communication by affecting multiple aspects of speech production, including rhythm and articulation [1]. As a result, Automatic Speech Recognition (ASR) systems trained on typical speech often struggle to process dysarthric speech accurately [2]. This creates a need for specialized ASR systems that accommodate the unique speech patterns of individuals with dysarthria. However, developing such systems is challenging for two main reasons. First, dysarthric speech deviates significantly from typical speech patterns, and exhibits substantial inter-speaker variability. Speech recognition models particularly struggle with the slower speaking rates common in dysarthria. Second, the scarcity of dysarthric speech data complicates development, as collecting speech samples can be physically demanding for individuals with dysarthria.

Recent studies have explored promising avenues such as the use of synthesis and conversion methods for data augmentation and dysarthric-to-healthy speech conversion [3, 4, 5]. Among these, unsupervised techniques leveraging Self-Supervised Learning (SSL) speech representations have demonstrated promise due to their zero-shot capabilities and low data requirements. In previous work, we presented an unsupervised Rhythm and Voice (RnV) conversion framework [6] that modifies dysarthric speech to resemble healthy speech, showing promise in improving ASR performance. Indeed, evaluations on the Torgo corpus [7] showed that rhythm conversion was particularly beneficial for speakers with more severe dysarthria, improving ASR performance on models trained solely on healthy speech. While RnV enables unsupervised and data-efficient dysarthric speech conversion, the rhythm modeling approach used was not specifically adapted to dysarthric speech, resulting in imprecise segmentation and rhythm modification. Furthermore, while ASR performance improved, it remained unsatisfactory, and the impact of training or adapting ASR models on converted speech was not explored.

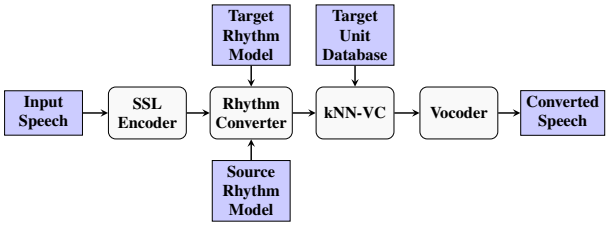


Figure 1: *RnV Framework Overview*

To address these limitations, this paper explores a syllable-based rhythm modeling approach aimed at improving segmentation and adaptation for dysarthric speech. We investigate the impact of training and adapting ASR models on converted speech by training an LF-MMI [8] model and fine-tuning Whisper [9], providing a comparative analysis of their performance.

2. Background

The RnV framework (Fig. 1) converts dysarthric speech into healthy speech in unsupervised fashion by leveraging properties of self-supervised speech representations [6]. Rhythm conversion is achieved through a modified version of Urhythmic [10], replacing soft units with discrete speech representations and extending the any-to-one conversion approach to any-to-any using a general-purpose vocoder. This method first segments speech into three types: silences, sonorants, and obstruents. This is done by clustering discrete units from a speech dataset into 100 centroids, which are then hierarchically grouped. The group with the greatest overlap with silences is labeled as Silence, the one most associated with voiced sections as Sonorants, and the remaining type as Obstruents. Segmentation is performed by comparing each discrete unit to the centroids and computing its log-probability of belonging to each class. A dynamic programming algorithm then merges consecutive units into longer segments, with a parameter γ controlling segment length. This segmentation enables the calculation of speaking rate and duration distributions for each speech type per speaker. These can be used to modify rhythm through time-stretching, either at the utterance level (global) or at the level of individual segments and speech types (fine-grained). Voice conversion, on the other hand, employs kNN-VC [11] to map the phonetic content of the source speaker to the closest matching units of a target speaker.

3. Methods

In this work, we extend the rhythm conversion module of the RnV framework by combining the unsupervised clustering-based method with syllable segmentation and modeling. We further train and adapt ASR models on the converted speech to assess more thoroughly whether conversion helps improve recognition performance.

3.1. Syllable-based rhythm modeling

Syllables are fundamental to rhythm modeling in speech. Indeed, analyzing syllable durations helps characterize a speaker's unique rhythmic patterns [12]. While time-aligned transcriptions are ideal for precise syllable segmentation and speaking rate calculation, alternative methods are needed when such transcriptions are unavailable. Prior work using the Urhythmic method estimates speaking rate by segmenting and counting sonorants per second, as sonorants serve as syllable nuclei [6]. However, obtaining precise sonorant segments proved challenging due to the unique characteristics of dysarthric speech. To address this, we aim to improve rhythm modeling by directly obtaining syllable segments in an unsupervised manner.

The importance of syllable-level analysis, particularly segmentation and feature extraction, has been demonstrated in recent research on dysarthric speech detection [13]. Indeed, such approaches mirror human auditory processing, where sound is initially decomposed into frequencies by the cochlea and subsequently segmented into syllables by the cortex. Inspired by this work, we adopt a segmentation approach which leverages theta oscillations and the sonority envelope [14]. This method segments syllables based on either syllable onsets (valleys in the envelope) or syllable nuclei (typically vowels, represented by peaks in the envelope). By extracting peak and valley timestamps, we can define syllable segments as either valley-to-valley or peak-to-peak intervals.

While effective, this method struggles with noisy speech, as noise can introduce false peaks and valleys. Given the presence of noise in the data which we will be using for evaluation, we incorporate a filtering step using the initial discrete unit clustering method. This allows us to ignore segments that fall outside speech regions. Specifically, we define speech segments as the combined obstruent and sonorant regions, while silence segments are classified as non-speech. This segmentation acts as a form of Voice Activity Detection (VAD), helping to distinguish speech from noise. In preliminary experiments, we observed that this approach was more robust than traditional VAD methods in identifying non-speech regions in the context of dysarthric speech. Any peak or valley within a non-speech region is discarded, improving segmentation accuracy. The segmentation steps are visualized in Figure 2.

After filtering, we focus on syllable nuclei (peaks), as they are less prone to false positives. By measuring peak-to-peak syllable durations for each speaker, we establish two rhythm modeling and conversion approaches:

Global: We calculate the syllables-per-second rate for each speaker to determine a global speaking rate. To convert a source speaker's utterance to a target speaker's rhythm, we time-stretch the discrete units at the utterance level using the ratio of their speaking rates.

Fine-Grained: Inspired by the fine-grained Urhythmic approach, we fit syllable durations to a gamma distribution, creating a speaker-specific duration model. To convert rhythm, we map each source syllable segment's duration to the target distribution using the Cumulative Distribution Function (CDF) and the Percent Point Function (PPF). This ensures that each segment's duration maintains its probability rank within the target distribution, preserving natural rhythm characteristics.

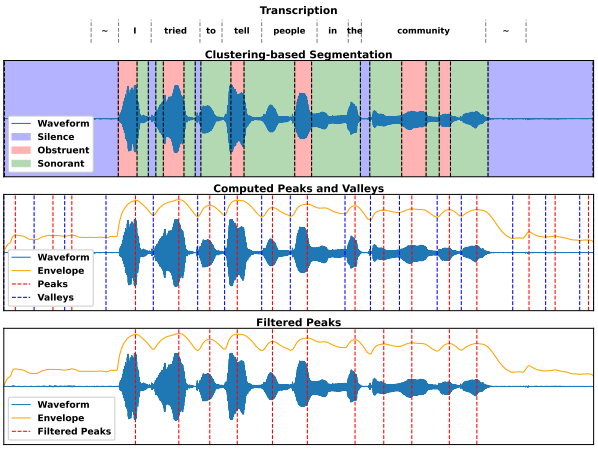


Figure 2: *Segmentation steps for Torgo speaker M02 pronouncing 'I tried to tell people in the community'.*

3.2. ASR adaptation

In this study, we investigate whether speech conversion enhances ASR performance by training and adapting models on the converted data. To achieve this, we explore two approaches: first, we train LF-MMI ASR models from scratch, and second, we fine-tune a pre-trained Whisper model originally trained on healthy speech. A detailed explanation of these methodologies is provided in the following section.

4. Experimental Setup

4.1. RnV implementation

We implement the framework similarly to [6]. We use the 6th layer of WavLM Large [15] as our speech representation, and reconstruct waveforms using a pre-trained HiFi-GAN V1 vocoder [16] checkpoint trained using the pre-matched paradigm from [11]. For the clustering-based segmentation, we use $\gamma = 3$. For kNN-VC, we find the $k = 8$ nearest units calculated using the cosine distance, and apply weighted averaging to the obtained units.

4.2. Datasets

Our evaluation is conducted using the Torgo corpus [7], which contains voice samples from 15 participants: 8 individuals with dysarthria (stemming from either Cerebral Palsy or Amyotrophic Lateral Sclerosis) and 7 control subjects. Speakers with dysarthria are classified into four severity levels: severe, moderately severe, moderate, and mild. The collected samples encompass 725 sentences and 2,340 isolated words. We use both the recordings from the head-mounted and array microphones. For the target speech in our conversion process, we use the LJSpeech database [17], which features 24 hours of single-speaker English audiobook narration. We process all audio samples by standardizing them to 16kHz, which is WavLM's expected input sampling rate, and normalizing volume levels to -20dB.

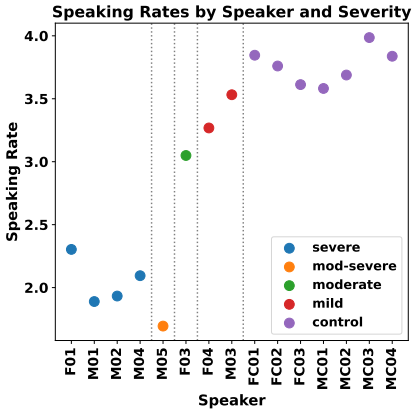


Figure 3: *Global speaking rates computed using the Syllable-based method for each Torgo speaker, categorized by severity.*

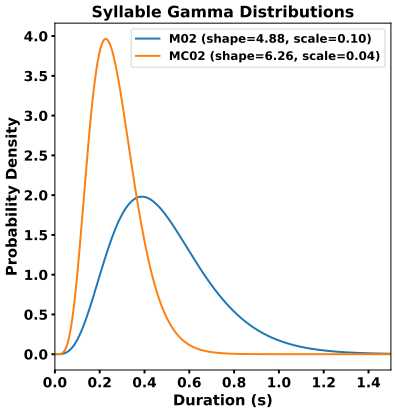


Figure 4: *Comparison of syllable gamma duration distributions for control speaker MC02 and dysarthric speaker M02.*

4.3. ASR experiments

We use a Leave-One-Speaker-Out approach, where to evaluate the performance on each speaker, we train/fine-tune each model on the data from all other speakers and test on the remaining speaker. To evaluate the impact of different conversion setups on ASR performance, we conduct experiments with two different, existing models. The proposed conversion methods could also be combined with other ASR models for potential better absolute performance, but we emphasized openly available implementations in this work.

LF-MMI: We train factorized time-delay neural network [18] acoustic models with the sequence-discriminative LF-MMI loss [8]. The models are trained in Kaldi [19] using the training recipe from [20], i.e. first training HMM-GMM ASR models and then using their alignments for LF-MMI training with speed perturbation (factors 0.9, 1.0, 1.1). The only difference is that we do not use i-vectors for simplicity. Also following [20], when decoding isolated words a grammar restricts the output to one of the possible options, for sentences a bigram language model is trained on all sentence data.

Whisper: We fine-tune the pre-trained Whisper base model [9], which comprises 74M parameters. It is pre-trained on a diverse multilingual dataset for generalization across languages, and incorporates a multitask learning framework which includes transcription, translation, and language identification. It employs a sequence-to-sequence approach, where audio inputs are converted into log-Mel spectrograms and processed by a convolutional encoder to extract features. These features are then passed through a transformer-based encoder, which captures long-range dependencies. A transformed decoder generates text tokens autoregressively, conditioned on the encoded audio representations. For fine-tuning, we use a batch size of 32, a learning rate $\alpha = 1e-5$, and early stopping with patience set to 5 epochs. There is no additional language model.

ASR performance is assessed using the word error rate (WER). First, we present the speaker-averaged WER results for Torgo speakers with dysarthria across all conversion setups. We then plot per-speaker WER results for selected setups for a more detailed analysis. Following the approach in [20], we report the results separately for isolated words and sentences, as each scenario presents distinct challenges.

Table 1: *WER Results averaged over Dysarthric Torgo speakers for all conversion setups*

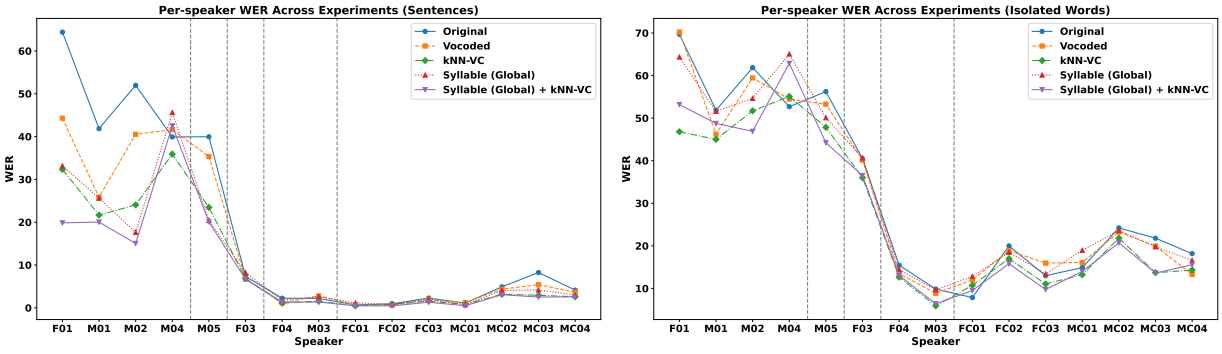
	LF-MMI		Fine-tuned Whisper	
Experiment	**Isolated**	**Sentences**	**Isolated**	**Sentences**
Original	44.8	31.2	48.35	29.62
Vocoded	43.3	24.8	50.66	32.38
kNN-VC	**37.6**	18.4	53.18	34.53
Urhythmic (Fine)	60.0	26.2	54.67	37.21
Urhythmic (Global)	42.6	20.6	49.51	30.09
Syllable (Fine)	46.5	20.6	60.01	33.94
Syllable (Global)	43.8	19.4	50.93	31.81
Urhythmic (Fine) + kNN-VC	52.1	17.9	56.26	34.10
Urhythmic (Global) + kNN-VC	38.4	**15.9**	51.19	39.49
Syllable (Fine) + kNN-VC	39.4	16.9	57.73	33.32
Syllable (Global) + kNN-VC	39.0	**15.9**	52.75	36.29

4.4. Conversion setups

For conversion, similarly to [6], we first train the Urhythmic segmenter on LJSpeech to obtain the 100 centroids, hierarchically group them into the three speech types, and perform segmentation as described in Section 2. Using both the Urhythmic and syllable-based methods, we compute global and fine-grained rhythm models for LJSpeech and each Torgo speaker, enabling the conversion of Torgo utterances to LJSpeech under different setups. We then evaluate and compare ASR performance across original, vocoded (encoded and decoded without modification), voice-converted, rhythm-converted (using each global and fine-grained method), and rhythm + voice-converted samples. In the next section, we refer to the rhythm conversion approaches as Syllable and Urhythmic, denoting the fine-grained and global methods as Fine and Global, respectively.

5. Results

Figure 3 presents the speaking rates calculated for each Torgo speaker using the syllable-based method. We can observe that speaking rates increase with lower severity levels as expected. Severe and moderately severe speakers exhibit rates around 2 syllables per second, while control speakers have a rate close to 4 syllables per second, which aligns with the typical average. Compared to speaking rates derived from counting sonorants per second, as reported in [6], the syllable-based method provides clearer separation between severity levels and more consistent rates for control speakers. Moreover, Figure 4 illustrates the syllable duration gamma distributions for control



(a) *Sentences* (b) *Isolated Words*

Figure 5: *Per-speaker WER results on Torgo using the LF-MMI model for different conversion setups.*

speaker MC02 and speaker with dysarthria M02. The probability density for the control speaker peaks just below 0.25 seconds, which would correspond to a rate of 4 syllables per second. In contrast, the distribution for the speaker with dysarthria is more variable, peaks at 0.4 seconds, and has a longer tail, with some syllable durations exceeding 1 second.

Table 1 presents the WER results averaged across dysarthric Torgo speakers for both the LF-MMI and fine-tuned Whisper-base models. For the Whisper-base models, the conversion methods do not yield improvements over fine-tuning on the original data, with WERs of approximately 30 for sentences and 50 for isolated words. In contrast, the LF-MMI model shows clear improvement. When trained on the original data, its performance is comparable to the Whisper models. Applying kNN-VC significantly enhances results, achieving the lowest WER for isolated words and reducing the WER to 18.4 for sentences. Rhythm conversion methods also improve performance: both global rhythm conversion approaches reduce WER to around 20 for sentences, while the syllable-based fine-grained method outperforms the Urhythmic-based fine-grained method and performs similarly to the global methods. Finally, combining rhythm conversion with kNN-VC further reduces WER, achieving the best overall sentence performance (15.9) when either global rhythm conversion method is used. For isolated words, this combination performs comparably to kNN-VC alone and outperforms rhythm conversion alone.

Figure 5 presents the WER results per speaker for different configurations. For sentences, the vocoded data interestingly outperforms the original data. Both kNN-VC and global syllable-based conversion show clear improvements over the original and vocoded data, particularly for speakers with more severe dysarthria. Combining these methods further reduces WER, except for speaker M04, where rhythm conversion has minimal impact. Performance for mild and control speakers on the other hand is largely unaffected or slightly improved by the conversion methods. For isolated words, kNN-VC provides the most significant improvements, while its combination with rhythm modeling yields additional gains in some cases, though the effect is less significant than for sentences.

6. Discussion and conclusions

The rhythm analysis and modeling results demonstrate that syllable-based segmentation is well-suited for dysarthric speech. The clear correlation between speaking rate and dysarthria severity supports this approach, as speaking rate increases with lower severity. Additionally, fitting a gamma distribution to each speaker's syllable durations provides a more detailed representation of individual rhythm characteristics. Beyond conversion, the ability to capture detailed rhythmic information could be helpful for diagnostic and assessment tools for dysarthric speech.

Dysarthric-to-healthy rhythm conversion proved particularly beneficial for the LF-MMI model, leading to noticeable WER reductions, especially for speakers with severe and moderately severe dysarthria. Global rhythm modeling methods performed best, as they rely on simple utterance-level time-stretching, minimizing errors and preventing artifacts. Further, these methods do not require highly precise speaking rate calculations, as long as the overall speaking rate is increased for highly severe cases. Among fine-grained rhythm modeling methods, the syllable-based approach was more effective than the fine-grained Urhythmic method. While the former achieved performance improvements comparable to global methods, the latter introduced artifacts due to imprecise segmentation, resulting in a drop in ASR performance. In addition, we observed that voice conversion using kNN-VC was just as effective in improving the LF-MMI model's performance. This is likely due to reducing the inter-speaker variations, which simplifies the task for the model. The best results were obtained by performing both rhythm and voice conversion, demonstrating that both techniques are complementary.

On the other hand, fine-tuned Whisper-base models did not benefit from rhythm or voice conversion, likely due to their extensive pre-training on over 680,000 hours of healthy speaker audio. This large-scale training enables Whisper to generalize well across speakers, making voice standardization unnecessary. Furthermore, once fine-tuned on dysarthric speech, the model's transformer architecture can adapt to slower speaking rates, reducing the need for rhythm conversion. Indeed, the fine-tuned models significantly outperform the pre-trained base model, which, as reported in [6], produced hallucinated outputs when not adapted to dysarthric speech. These hallucinations are caused by slower speaking rates and resulted in excessively high WER values for speakers with severe dysarthria. In contrast, the LF-MMI model benefited from rhythm and voice conversion as it was trained from scratch. Converting all Torgo data to a single speaker reduced inter-speaker variations, simplifying the training process and enhancing test-time performance on similarly converted data. Future work could focus on identifying distinct syllable groups to enable more detailed rhythm modeling.

7. Acknowledgements

This work was partially supported by the Swiss National Science Foundation (SNSF) through the project "Pathological Speech Synthesis (PaSS)" (grant agreement no. 219726), by the SNSF through the Bridge Discovery project "Emotion in the loop - a step towards a comprehensive closed-loop deep brain stimulation in Parkinson's disease (EMIL)" (grant agreement no. 40B2–0_194794), and by the Innosuisse through the flagship project "Inclusive Information and Communication Technologies (IICT)" (grant agreement no. PFFS-21-47).

8. References

[1] J. R. Duffy, *Motor Speech Disorders*, 3rd ed. Mosby, 2012.

[2] M. Moore, H. Venkateswara, and S. Panchanathan, "Whistle-blowing ASRs: Evaluating the Need for More Inclusive Speech Recognition Systems," in *Proc. Interspeech*, 2018, pp. 466–470.

[3] M. Soleymanpour, M. T. Johnson, R. Soleymanpour, and J. Berry, "Synthesizing Dysarthric Speech Using Multi-Speaker TTS For Dysarthric Speech Recognition," in *Proc. ICASSP*, May 2022, pp. 7382–7386.

[4] E. Hermann and M. Magimai.-Doss, "Few-shot Dysarthric Speech Recognition with Text-to-Speech Data Augmentation," in *Proc. Interspeech*, Aug. 2023, pp. 156–160.

[5] W.-Z. Leung, M. Cross, A. Ragni, and S. Goetze, "Training Data Augmentation for Dysarthric Automatic Speech Recognition by Text-to-Dysarthric-Speech Synthesis," in *Proc. Interspeech*, 2024, pp. 2494–2498.

[6] K. E. Hajal, E. Hermann, A. Kulkarni, and M. Magimai.-Doss, "Unsupervised rhythm and voice conversion of dysarthric to healthy speech for ASR," in *Proc. Workshop on Speech Pathology Analysis and DEtection (SPADE) at ICASSP*, 2025.

[7] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," in *Proc. LREC*, vol. 46, no. 4, 2012, pp. 523–541.

[8] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI," in *Proc. Interspeech*, 2016, pp. 2751–2755.

[9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023.

[10] B. van Niekerk, M.-A. Carbonneau, and H. Kamper, "Rhythm modeling for voice conversion," *IEEE Signal Processing Letters*, vol. 30, pp. 1297–1301, 2023.

[11] M. Baas, B. van Niekerk, and H. Kamper, "Voice conversion with just nearest neighbors," in *Proc. Interspeech*, 2023, pp. 2053–2057.

[12] T. Pfau and G. Ruske, "Estimating the speaking rate by vowel detection," in *Proc. ICASSP*, vol. 2, 1998, pp. 945–948 vol.2.

[13] S. Hovsepyan and M. Magimai.-Doss, "Syllable level features for Parkinson's disease detection from speech," in *Proc. ICASSP*, 2024, pp. 11 416–11 420.

[14] O. Räsänen, G. Doyle, and M. C. Frank, "Pre-linguistic segmentation of speech into syllable-like units," *Cognition*, vol. 171, pp. 130–150, 2018.

[15] 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*, pp. 1505–1518, 2022.

[16] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, 2020, pp. 17 022–17 033.

[17] K. Ito and L. Johnson, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[18] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in *Proc. Interspeech*, 2018, pp. 3743–3747.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi Speech Recognition Toolkit," in *Proc. ASRU*, 2011.

[20] E. Hermann and M. Magimai.-Doss, "Dysarthric speech recognition with lattice-free MMI," in *Proc. ICASSP*, 2020, pp. 6109–6113.