EA-VAE: Learning to Reconstruct Dysarthric Speech via Variational Autoencoder with Encoding Alignment

Daipeng Zhang[1], **Wenhuan Lu**[1,2,3*], **Xianghu Yue**[2], **Hongcheng Zhang**[2], **Jianguo Wei**[1,2,3]

[1]School of New Media and Communication, Tianjin University, Tianjin, China
[2]College of Intelligence and Computing, Tianjin University, Tianjin, China
[3]School of Intelligence Science and Engineering, Qinghai Minzu University, Xining, China
zhangdaipeng@tju.edu.cn, wenhuan@tju.edu.cn, yuexianghu@tju.edu.cn,
zhanghongcheng@tju.edu.cn, jianguo@tju.edu.cn

Abstract

Dysarthric speech reconstruction (DSR) aims to enhance the intelligibility of dysarthric speech. Compared with normal speech, the dysarthric speech is characterized by its pathological features, including discontinuous pronunciation, slow speech, hoarseness, and improper pauses. Significant disparities in the feature space between normal and dysarthric speech may result in suboptimal speech reconstruction, thereby degrading speech intelligibility. To enhance the reconstruction ability of speech feature spaces, this paper proposes a DSR model named the Encoding-Aligned Variational Autoencoder (EA-VAE). By incorporating alignment modules of frame-level embedding features, prior distributions, and duration into the encoder of the VAE, the model explicitly aligns the dysarthric speech encoding with a representation of the parallel normal speech. A shared decoder is then used to generate speech with improved intelligibility. Experimental results on the UASpeech benchmark confirm that EA-VAE achieves state-of-the-art performance, with a 31.7% relative word error rate reduction and the highest subjective MOS score (4.48), thoroughly validating the effectiveness and advancements of the proposed method in dysarthric speech reconstruction.

Introduction

Dysarthria refers to a group of motor speech disorders caused by deficits in neuromuscular control, commonly associated with a range of neurological conditions such as Parkinson's Disease, Amyotrophic Lateral Sclerosis, and Cerebral Palsy (Almadhor et al. 2023). Individuals with moderate to severe dysarthria experience impaired muscular coordination, which is essential for proper articulation. This results in symptoms such as indistinct speech, abnormal speech rates, and fluctuations in prosody and vocal intensity. These pathological features significantly impact speech clarity and intelligibility (Almadhor et al. 2023; Zheng et al. 2023; Ueno, Lee, and Kawahara 2024).

Dysarthric speech reconstruction (DSR) techniques (Zheng et al. 2023; Wang et al. 2024) aim to enhance speech intelligibility by transforming impaired speech signals into acoustically normal speech (Aihara, Takiguchi, and Ariki 2017; Chen et al. 2024a). A major challenge in DSR lies in the substantial acoustic patterns differences between dysarthric and normal speech (Lee, Littlejohn, and Simmons 2017; Wang et al. 2024). Dysarthric speech presents a range of pathological features (Shahamiri, Lal, and Shah 2023; Yu, Su, and Qian 2023; Zhang et al. 2025), including discontinuous pronunciation, uncontrolled volume, slow speech, improper pauses, explosive pronunciation, hoarseness, and air-flow noise.

Current DSR approaches can be categorized into rule-based and statistical methods. Rule-based methods rely on handcrafted rules derived from expert knowledge, typically targeting phoneme error correction or modifications to temporal and spectral features to enhance speech intelligibility (Rudzicz 2011; Kumar and Kumar 2016). In contrast, statistical methods focus on automatically mapping dysarthric speech features to normal speech. Notable techniques include Gaussian Mixture Models (GMM) (Kain et al. 2007), Non-negative Matrix Factorization (NMF) (Aihara et al. 2012, 2013), Partial Least Squares (PLS) (Aihara, Takiguchi, and Ariki 2017), and deep learning-based approaches, i.e., two-stage (Matsubara et al. 2021; Liu et al. 2024; Wang et al. 2024) and end-to-end methods (Imai et al. 2020; Prananta et al. 2022; Purohit et al. 2020; Wang et al. 2020, 2022, 2024).

Two-stage models typically divide the task into two components: a content extraction stage based on voice conversion (VC) or automatic speech recognition (ASR) (Chen et al. 2024c,b), and a speech synthesis stage, based on text-to-speech (TTS) or vocoding techniques (Matsubara et al. 2021; Liu et al. 2024). VC-based methods often lack explicit linguistic supervision, which may lead to content distortion. ASR-based methods heavily rely on the performance of the ASR system, which is compromised by the unclear articulation, abnormal prosody, and limited availability of large, annotated dysarthric datasets. These limitations often result in poor global consistency and error propagation, ultimately degrading the quality of the synthesized speech in the second stage. End-to-end methods address DSR as a cross-domain feature transformation problem by learning a direct mapping from dysarthric to normal speech features (Imai et al. 2020; Prananta et al. 2022; Purohit et al. 2020). These models in-

*Corresponding author.

corporate dedicated encoders for content extraction, prosody correction, and speaker identity preservation, followed by a decoder and vocoder for waveform reconstruction (Wang et al. 2020, 2022, 2024). To improve linguistic representation learning, end-to-end systems employ auxiliary strategies such as cross-modal knowledge distillation (Wang et al. 2020), phonetic posteriorgram features (Wang et al. 2022), and phoneme recognition-based modeling (Biadsy et al. 2019; Chen et al. 2021; Doshi et al. 2021). Despite their effectiveness, these methods typically require multiple specialized encoders, complex auxiliary learning schemes, and annotated data such as phoneme or text labels, resulting in a complicated training pipeline with multi-stage optimization.

To solve the aforementioned issues, we propose a simple yet effective framework, namely Encoding-Aligned Variational Autoencoder (EA-VAE), that only leverages a variational autoencoder (VAE) to transform dysarthric speech into intelligible and natural-sounding speech. EA-VAE builds upon the VITS (Kim, Kong, and Son 2021) framework to leverage its high-quality speech synthesis capabilities while disentangling content information without relying on textual annotations. The model comprises three key components. First, the Embedding Alignment Module (EAM) projects frame-level dysarthric speech embeddings to the corresponding embeddings of normal speech, mitigating discrepancies in acoustic patterns. Second, the Distribution Alignment Module (DAM) employs KL divergence to align the frame-level latent distributions of dysarthric and normal speech, thereby reducing distributional shift. Third, to address mismatches in pronunciation duration, a Duration Alignment Predictor (DAP) adjusts temporal patterns to improve the fluency and intelligibility of the reconstructed speech. The main contributions are summarized as follows:

- We present the EA-VAE, a streamlined yet effective framework for DSR that relies solely on a VAE.
- We introduce the EAM module to align the frame-level dysarthric and normal speech embeddings by modeling both local dynamics and global dependencies.
- To minimize the distributional shift between dysarthric and normal speech, we propose the DAM module, which aligns prior distributions inferred from dysarthric speech with the parallel normal speech.
- We develop a DAP to capture duration patterns and align dysarthric speech durations with natural prosody.
- Extensive experiments on the UASpeech benchmark demonstrate that EA-VAE achieves state-of-the-art performance, with comprehensive analysis validating its effectiveness and improvements over existing methods. Our demo has been publicly released [1].

Related Work

Dysarthric speech reconstruction (DSR) aims to enhance intelligibility and naturalness, and can be divided into two main categories: two-stage and end-to-end methods.

[1]https://sailbulider.github.io/EA-VAE/.

Two-stage DSR Two-stage methods have shown promise in improving the intelligibility of dysarthric speech by decomposing the task into sequential modules. One representative approach combines Transformer-based TTS with CycleVAE-based voice conversion, followed by LPCNet vocoding, to generate intelligible speech while preserving speaker individuality by transferring speaker characteristics to the synthesized output (Matsubara et al. 2021). Another method adopts a two-stage zero-shot voice conversion pipeline, where dysarthric speech is first repaired using a gender-constrained KNN-based retrieval mechanism and then refined using so-vits-svc to enhance timbre and naturalness (Liu et al. 2024). Additionally, the baseline proposed in (Wang et al. 2024) employs a modular ASR-TTS architecture that integrates HuBERT-CTC for content extraction and Tacotron 2 for synthesis, followed by HiFi-GAN for vocoding (Shen et al. 2018; Kong, Kim, and Bae 2020).

End-to-end DSR End-to-end methods aim to directly convert dysarthric speech into intelligible and natural-sounding speech. MMSE-DiscoGAN (Purohit et al. 2020) outperforms conventional DNNs by learning direct feature mappings between dysarthric and normal speech. CycleGAN-based models (Imai et al. 2020; Yang and Chung 2020) support non-parallel voice conversion and reduce ASR error rates. StarGAN variants (Zheng et al. 2023; Mehrez, Chaiani, and Selouani 2024) further enhance clarity and naturalness, particularly when combined with data augmentation techniques. E-DGAN (Chu et al. 2023) adopts an encoder-decoder architecture to generalize across pathological speech conditions while preserving speaker individuality. A cross-modal knowledge distillation model (Wang et al. 2020) leverages a TTS-trained encoder as a teacher to guide a speech encoder for direct dysarthric-to-normal speech conversion. To improve speaker identity preservation, adversarial speaker adaptation (Wang et al. 2022) finetunes the speaker encoder while regularizing output distributions. UNIT-DSR (Wang et al. 2024) simplifies the pipeline by using discrete speech units and HuBERT-based normalization, achieving competitive reconstruction quality with improved robustness and efficiency. The Parrotron series (Biadsy et al. 2019; Chen et al. 2021; Doshi et al. 2021) introduces spectrogram-to-spectrogram models based on encoder-decoder architectures combined with vocoders.

Two-stage methods rely on complex content extraction and speech synthesis processes, while end-to-end methods use complex training strategies. Unlike the aforementioned work, we simplify the process by incorporating alignment modules for frame-level embedding features, prior distributions, and duration directly into the VAE encoder, avoiding the need for specialized encoders, auxiliary strategies, or text annotations. This simplification reduces the accumulation of errors and enhances the consistency and naturalness of speech reconstruction.

Methodology

Problem Definition

The original inputs for DSR include dysarthric speech audio x and parallel normal speech audio y. The goal of this

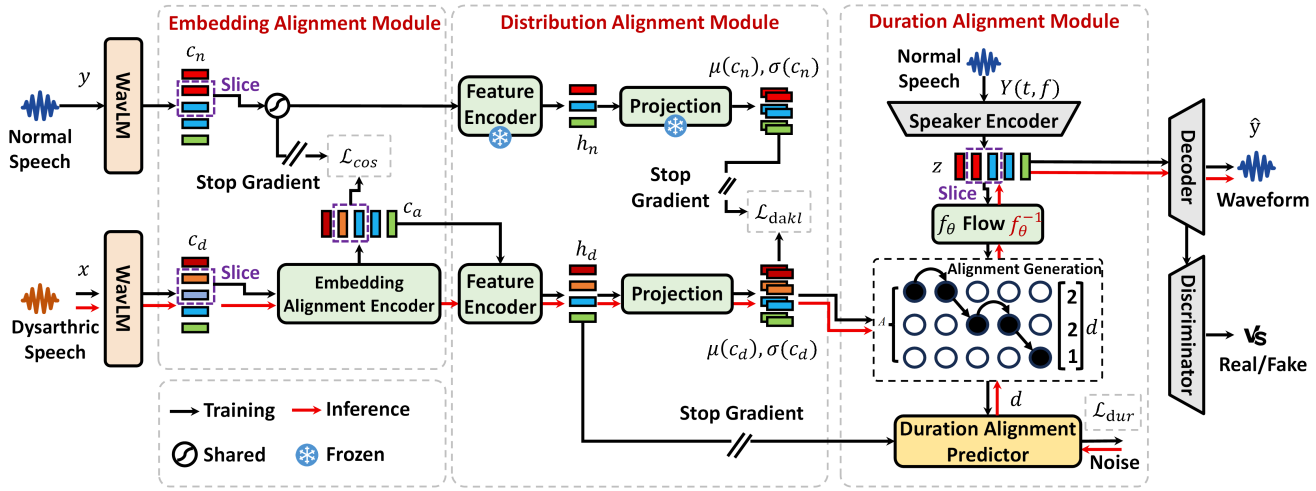

Figure 1: The overall framework of our proposed EA-VAE, starting with an Embedding Alignment Module (EAM) that aligns frame-level embeddings between dysarthric and normal speech, followed by a Distribution Alignment Module (DAM) that aligns the prior distributions inferred from dysarthric speech with those of normal speech, finally a Duration Alignment Predictor (DAP) which models duration patterns, ensuring dysarthric speech aligns with natural prosody.

task is to learn a mapping function that transforms x into an intelligible and natural-sounding normal speech $\hat{y}$.

Model Overview

EA-VAE can be interpreted as a conditional VAE (Kim, Kong, and Son 2021), aiming to maximize the evidence lower bound (ELBO), of the intractable marginal log-likelihood of the data $\log p(\hat{y} \mid c)$:

$$\log p(\hat{y} \mid c) \geq \mathbb{E}_{q(z|y)} \left[\log p(\hat{y} \mid z) - \log \frac{q(z \mid y)}{p(z \mid c)} \right], \quad (1)$$

where $p(z \mid c)$ represents the prior distribution of the latent variable z given condition c, while $p(\hat{y} \mid z)$ serves as the likelihood function for $\hat{y}$, and $q(z \mid y)$ is the approximate posterior distribution.

The condition vector c represents the frame-level embedding fratures of the dysarthric speech x, which are extracted using the pre-trained WavLM (Chen et al. 2022) model $f_{\text{WavLM}}(\cdot)$:

$$c = f_{\text{WavLM}}(x), c \in \mathbb{R}^{T \times d}, \quad (2)$$

where T denotes the number of frames and d is the embedding dimension.

The training objective is to minimize the negative ELBO, which can be interpreted as the combination of the reconstruction loss, $-\log p(\hat{y} \mid z)$, and the KL divergence, $\text{KL}[q(z|y) \| p(z|c)]$, where $z \sim q(z \mid y)$.

Given the significant acoustic pattern differences between dysarthric and normal speech, we designed several modules to enhance the strength of the prior distribution $p(z|c)$. The overall processing pipeline of the proposed EA-VAE is shown in Figure 1. We begin by designing the Embedding Alignment Module (EAM) which aims to align the frame-level embeddings of dysarthric speech with those of parallel normal speech. To ensure the consistency in the latent space, we further introduce a Distribution Alignment Module (DAM), which aligns the prior distributions inferred from dysarthric speech with those of normal speech. Since dysarthric speech exhibits duration abnormalities, we finally develop a Duration Alignment Predictor (DAP) to improve duration modeling.

Embedding Alignment Module

The Embedding Alignment Module (EAM) aligns the frame-level embeddings of dysarthric speech $c_{\text{d}} = f_{\text{WavLM}}(x)$ to those of parallel normal speech $c_{\text{n}} = f_{\text{WavLM}}(x)$. The EAM initially extracts a transformed representation c_{a} from the original embedding c_{d} via the Embedding Alignment Encoder. In addition, alignment is achieved by maximizing the cosine similarity between c_{a} and c_{d}, encouraging semantic consistency.

We design an Embedding Alignment Encoder using Mamba blocks used in LNVAE (Zhang et al. 2025) and Transformer blocks (Vaswani et al. 2017). The Mamba blocks capture local temporal dynamics such as long-range dependence and non-stationarity (Zhang et al. 2025), while the Transformer module models global dependencies through self-attention.

To avoid interference from non-informative frames, this encoder introduces masking encoding matrix $P \in \mathbb{R}^{|c_{\text{d},t}|}$ in the Mamba blocks, allowing them to handle irregular or missing segments by ignoring masked time steps, thereby enhancing the robustness and expressiveness of the learned mapping c_{a}. The Mamba blocks' output is expressed as:

$$c_{\text{a,t}} = Cs_{\text{t}}, s_{\text{t}} = As_{\text{t-1}} + B(c_{\text{d,t}} \cdot P), \quad (3)$$

where A, B, and C denote the learned parameter matrices

of the Mamba blocks, the hidden state s_{t} at time step t is updated based on the previous hidden state $s_{\mathrm{t\text{-}1}}$ at time step $t-1$.

Similarly, this Transformer blocks introduce a masking mechanism, represented as:

$$\begin{aligned}&\mathrm{Attention}(Q,K,V)=\\&\sum \mathrm{Softmax}\left(\frac{QW_{\mathrm{Q}}\left(KW_{\mathrm{K}}\right)^{T}+QW_{\mathrm{Q}}\left(P^{K}\right)^{T}}{\sqrt{d_{\mathrm{z}}}}\right)\left(VW_{\mathrm{V}}\right),\end{aligned} \tag{4}$$

where, $W_{\mathrm{Q}}, W_{\mathrm{K}}, W_{\mathrm{V}}$ are the attention parameters.

To optimize the Embedding Alignment Encoder, we compute the similarity between c_{a} and c_{n} using cosine similarity. The loss function can be defined as:

$$\mathcal{L}_{\mathrm{cos}}=1-\frac{c_{\mathrm{a}}\cdot c_{\mathrm{n}}}{\|c_{\mathrm{a}}\|\,\|c_{\mathrm{n}}\|}. \tag{5}$$

Distribution Alignment Module

The Distribution Alignment Module (DAM) is designed to align the prior distribution of dysarthric speech $p(z \mid c_{\mathrm{d}})$ with the prior distribution of normal speech $p(z \mid c_{\mathrm{n}})$. The DAM begins by using a Feature Encoder to extract latent representations h_{d} from the embeddings c_{d}. Furthermore, following the parameterization process of Kim et al. (Kim, Kong, and Son 2021), the prior distribution $p(z \mid c_{\mathrm{d}})$ and $p(z \mid c_{\mathrm{n}})$ is obtained by parameterizing h_{d} and h_{n} with a factorized normal distribution. Finally, the DAM introduce a Distribution Alignment Mechanism, which guides the model to learn $p(z \mid c_{\mathrm{d}})$ consistent with $p(z \mid c_{\mathrm{n}})$. $p(z \mid c_{\mathrm{d}})$ and $p(z \mid c_{\mathrm{n}})$ can be represented as:

$$\begin{aligned}p(z \mid c_{\mathrm{d}})&=\mathcal{N}\left(z;\mu(c_{\mathrm{d}}),\sigma(c_{\mathrm{d}})\right),\\p(z \mid c_{\mathrm{n}})&=\mathcal{N}\left(z;\mu(c_{\mathrm{n}}),\sigma(c_{\mathrm{n}})\right),\end{aligned} \tag{6}$$

where $\mathcal{N}$ represents the factorized normal distribution, μ, σ represent mean and variance. Specially, the prior distribution of normal speech parameters is obtained through normal speech pre-training.

To encode the frame-level embedding features, we adopt the non-causal WaveNet residual blocks used in WaveGlow (Prenger, Valle, and Catanzaro 2019) and Glow-TTS (Kim et al. 2020) as our Feature Encoder. These blocks, composed of dilated convolutions with gated activations and skip connections, enable effective modeling of temporal dependencies in the embedding feature space. The output layer projects the encoded features to the mean and variance of a Gaussian posterior.

To reduce the distributional shift between dysarthric and normal speech, we introduce a Distribution Alignment Mechanism. Specifically, the prior distribution inferred from dysarthric speech is regularized by computing the KL divergence between the dysarthric prior $p(z \mid c_{\mathrm{d}})$ and the normal prior $p(z \mid c_{\mathrm{n}})$, expressed as:

$$\mathcal{L}_{\mathrm{dakl}}=KL\big(p(z \mid c_{\mathrm{d}}) \,||\, p(z \mid c_{\mathrm{n}})\big). \tag{7}$$

In addition, to reduce KL divergence and improve alignment stability, we make $T_{\mathrm{d}} \geq T_{\mathrm{n}}$ that ensures more reliable estimation of the latent distribution. T_{d} and T_{n} denote the number of frames for dysarthric and normal speech, respectively.

Let define $T_{\mathrm{d}} < T_{\mathrm{n}}$, The KL divergence is expressed as:

$$\begin{aligned}&\mathrm{KL}\left(\mathcal{N}\big(\mu(c_{\mathrm{d}}),\sigma(c_{\mathrm{d}})\big)\,\big\|\,\mathcal{N}\big(\mu(c_{\mathrm{n}}),\sigma(c_{\mathrm{n}})\big)\right)\\&=\frac{1}{2}\left[\log\left(\frac{\sigma(c_{\mathrm{n}})}{\sigma(c_{\mathrm{d}})}\right)+\frac{\sigma(c_{\mathrm{d}})}{\sigma(c_{\mathrm{n}})}+\frac{\big(\mu(c_{\mathrm{d}})-\mu(c_{\mathrm{n}})\big)^{2}}{\sigma(c_{\mathrm{n}})}-1\right].\end{aligned} \tag{8}$$

As shown in (8), due to the shared linguistic content, the two distributions exhibit similar means, that is $\mu(c_{\mathrm{d}}) \approx \mu(c_{\mathrm{n}})$, with their primary differences reflected in the variances. Then The KL divergence can be approximated as:

$$\begin{aligned}&\mathrm{KL}\left(\mathcal{N}\big(\mu(c_{\mathrm{d}}),\sigma(c_{\mathrm{d}})\big)\,\big\|\,\mathcal{N}\big(\mu(c_{\mathrm{n}}),\sigma(c_{\mathrm{n}})\big)\right)\\&=\frac{1}{2}\left[\log\left(\frac{\sigma(c_{\mathrm{n}})}{\sigma(c_{\mathrm{d}})}\right)+\frac{\sigma(c_{\mathrm{d}})}{\sigma(c_{\mathrm{n}})}-1\right].\end{aligned} \tag{9}$$

When $T_{\mathrm{d}} < T_{\mathrm{n}}$, then $\sigma(c_{\mathrm{n}}) > \sigma(c_{\mathrm{d}})$, the KL divergence is greater than zero and monotonically increasing, resulting in higher KL divergence and degraded alignment stability.

To increase the temporal resolution of dysarthric speech embeddings, we apply a time-stretching operation at the waveform level before feature extraction. Specifically, the dysarthric waveform is temporally stretched by a factor $\beta > 1$, resulting in a slower version of the original speech. The stretched audio is then passed through WavLM to obtain the new frame-level embeddings:

$$c_{\mathrm{d}}=f_{\mathrm{WavLM}}(\mathrm{Stretch}(x,\beta)), \quad \beta>1. \tag{10}$$

This process increases the number of frames T_{d} such that $T_{\mathrm{d}} \geq T_{\mathrm{n}}$.

Duration Alignment Module

The Duration Alignment Module starts by using a Duration Alignment Predictor (DAP) to learns duration d patterns from h_{d}, which are modified to reflect natural duration patterns. Then the Duration Alignment Module uses a Speaker Encoder (Kim, Kong, and Son 2021) and a normalizing Flow (Kim, Kong, and Son 2021) f_θ to get to extract the frame-level latent representation z, and then using the Monotonic Alignment Search (MAS) (Kim et al. 2020) align h_{d} with z. Specifically, an inverse flow f_θ^{-1} is applied during inference.

The DAP incorporates the Mamba module (Gu and Dao 2023; Zhang et al. 2025) into the existing framework, in order to alleviate the inherent randomness in the synthesis process. Specifically, the module employs the Mamba module to recode the duration d, given as:

$$\widehat{d}=AdP+Bh_{\mathrm{d}}+Cm_{\mathrm{t}}, \tag{11}$$

where $d, \widehat{d} \in \mathbb{R}^{n\times m}$, n, m, P denote the time resolution, dimension, and duration mask, respectively.

Further, define two random variables $u \in \mathbb{R}^{n\times m}, 0 \leq u \leq 1$ and $v \in \mathbb{R}^{1\times m}$, and take $\widehat{d}-u$ to be a sequence of positive real numbers, which the proposed method concatenate to denote the higher dimensional latent representations

by v and $\widehat{d}.$ such that u, v satisfy the posterior distribution $q_\phi(u, v \mid \widehat{d}, h_\mathrm{d})$. The resulting objective represents a variational lower bound on the log-likelihood of phoneme durations:

$$\log p_\theta(\widehat{d} \mid h_\mathrm{d}) \geq \mathbb{E}_{\widehat{d}_\phi(u,v|\widehat{d},h_\mathrm{d})}\left[\log \frac{p_\theta\left(\widehat{d}-u, v \mid h_\mathrm{d}\right)}{q_\phi(u, v \mid \widehat{d}, h_\mathrm{d})}\right]. \tag{12}$$

The loss permutation function is ELBO:

$$\mathcal{L}_\mathrm{dur} = \log p_\theta(\widehat{d} \mid h_\mathrm{d}) - \mathbb{E}_{q_\phi(u,v|\widehat{d},h_\mathrm{d})}\left[\log \frac{p_\theta(\widehat{d}-u, v \mid h_\mathrm{d})}{q_\phi(u, v \mid \widehat{d}, h_\mathrm{d})}\right]. \tag{13}$$

The Speaker Encoder and Flow module follow the design architecture of VITS (Kim, Kong, and Son 2021). Notably, the speaker encoder employs the short-time Fourier transform (STFT) to convert the normal waveform $s(t)$ into linear spectrograms $Y(t, f)$, as given:

$$Y(t, f) = \sum_{n=0}^{N-1} s[n] \cdot w[n-t] \cdot e^{-j\frac{2\pi}{N}fn}, \tag{14}$$

where $s[n]$ denotes the discrete value of the original dysarthric speech signal at the n-th sampling point. $w[n-t]$ refers to the window function applied at time t to extract the local signal. N represents the window length used in the Fourier transform.

Then the speaker encoder is used to compress $Y(t, f)$ into a frame-level representation z, from which the posterior distribution $q(z \mid y)$, given as:

$$z \sim q(z \mid y) = \mathcal{N}\left(z; \mu(Y(t, f)), \sigma(Y(t, f))\right), \tag{15}$$

where $\mathcal{N}$ represents the factorized normal distribution, μ, σ represent mean and variance.

Decoder and Discriminator

The Decoder, and Discriminator follow the design architecture of VITS (Kim, Kong, and Son 2021). The Discriminator D is used to distinguish the output produced by the Decoder G from the final generated waveform with the following loss function:

$$\begin{aligned} \mathcal{L}_\mathrm{adv}(D) &= \mathbb{E}_{(y,z)}\left[(D(y)-1)^2 + (D(G(z)))^2\right], \\ \mathcal{L}_\mathrm{adv}(G) &= \mathbb{E}_z\left[(D(G(z))-1)^2\right]. \end{aligned} \tag{16}$$

In addition, we use the feature loss,which is measured by the L1 distance between the real and fake samples of the intermediate features in each layer of the Discriminator, given as:

$$\mathcal{L}_\mathrm{fm}(G) = \mathbb{E}_{(y,z)}\left[\sum_l \left\|D^l(y) - D^l(G(z))\right\|_1\right], \tag{17}$$

where l denotes the number of layers of the Discriminator, $D^l(\cdot)$ denotes the l-th feature.

Final Loss

The overall training loss of the EA-VAE is denoted as:

$$\begin{aligned} \mathcal{L}_\mathrm{total} = \mathcal{L}_\mathrm{dur} + \mathcal{L}_\mathrm{recon} + \mathcal{L}_\mathrm{cos} + \mathcal{L}_\mathrm{kl} + \mathcal{L}_\mathrm{dakl} + \\ \mathcal{L}_\mathrm{adv}(G) + \mathcal{L}_\mathrm{fm}(G). \end{aligned} \tag{18}$$

Experiments

Datasets

We conduct experiments using the UASpeech corpus (Kim et al. 2008), which is widely regarded as one of the largest benchmark dataset for English dysarthric speech reconstruction. The corpus includes recordings from 15 dysarthric and 13 normal speakers, each contributing three blocks (B1–B3) with 765 isolated words, including 465 distinct words (three repetitions of 155 words for recognizer training and testing) and 300 uncommon words designed to enhance phone-sequence diversity. The training set includes recordings from blocks B1 and B3, while the evaluation uses recordings from block B2. Following (Wang et al. 2020, 2022, 2024), Four patients with different severity levels, F02 (Low), M07 (Low), F04 (Mid), and M05 (Mid) are selected for subjective and objective evaluation.

Training Details

We first apply waveform-level time-stretching to ensure $T_\mathrm{d} \geq T_\mathrm{n}$, and extract frame-level embeddings using the pretrained Large WavLM model[2]. Our training follows a three-stage procedure: (1) replacing dysarthric embeddings c_d with normal-speech embeddings c_n and pretraining on VCTK(Yamagishi et al. 2019); (2) fine-tuning this model on the CF02 subset of UASpeech (B1/B3 for training, B2 for testing; 5,355 utterances, 2.85 hours); and (3) freezing the normal-speech mapping channel and further adapting the model using c_d from 13 UASpeech speakers (B1/B3 for training, B2 for testing; 73,675 utterances, 66 hours). All stages share an identical training framework.

Evaluation Metrics

We conduct experiments on two evaluation metrics: The 5-scale mean opinion score (MOS) test and word error rate (WER). 20 participants are recruited to assess 10 randomly selected words from the B2 dataset. Following (Wang et al. 2020, 2022, 2024), the evaluation exclusively focuses on the semantic content similarity between the reconstructed speech and the corresponding reference speech of CF02 with explicit instructions to disregard speaker identity and background noise. In addition, the publicly available ASR system Jasper (Li et al. 2019) with greedy decoding, is employed to compute the WER on the B2 set.

Spectrogram Analysis

Figure 2 shows spectrograms from each model for the word "delete", using utterances from speakers F04 (Mid) and M07 (Low). Compared to the reference, the original dysarthric

[2]https://github.com/microsoft/unilm/tree/master/wavlm.

Method	Low ↓		Mid ↓		Average / △ ↓
	F02 / △	M07 / △	F04 / △	M05 / △	
Original (Wang et al. 2020)	95.9 / –	95.6 / –	81.7 / –	91.0 / –	91.05 / –
ASR-TTS (Wang et al. 2024)	81.6 / -14.9%	70.0 / -26.8%	75.4 / -7.7%	74.2 / -18.5%	75.3 / -17.3%
E2E-DSR (Wang et al. 2020)	72.0 / -24.9%	73.1 / -23.5%	69.3 / -15.2%	69.8 / -23.3%	71.05 / -22.0%
ASA-DSR (Wang et al. 2022)	65.8 / -31.4%	62.7 / -34.4%	65.6 / -19.7%	62.5 / -31.3%	64.15 / -29.6%
Unit-DSR (Wang et al. 2024)	68.3 / -28.8%	62.1 / -35.0%	65.5 / -19.8%	64.4 / -29.2%	65.08 / -28.5%
EA-VAE	**63.81 / -33.5%**	**60.64 / -36.6%**	**62.60 / -23.4%**	**61.72 / -32.2%**	**62.19 / -31.7%**

Table 1: Comparison of Word Error Rate (WER ↓) using different DSR methods on the UASpeech. △ refers to the percentage decrease in WER compared to original dysarthric speech.

Method	Low ↑		Mid ↑		Average ↑
	F02	M07	F04	M05	
Ground Truth	4.92	4.88	4.85	4.80	4.86
Origin	2.20	2.10	3.21	3.36	2.72
ASA-DSR	3.64	3.36	3.52	4.17	3.67
ASR-TTS	3.65	4.44	4.41	3.92	4.11
E2E-DSR	3.89	3.92	3.52	4.00	3.83
Unit-DSR	**4.42**	4.32	4.31	4.50	4.39
EA-VAE	4.02	**4.54**	**4.69**	**4.65**	**4.48**

Table 2: The 5-scale MOS test scores (↑) of "F02", "M07", "F04", and "M05" on the UASpeech for content similarity with mean scores and the 95% confidence intervals.

Method	Low ↑		Mid ↑		Average ↑
	F02	M07	F04	M05	
Ground Truth	4.92	4.88	4.85	4.80	4.86
Origin	2.20	2.10	3.21	3.36	2.72
w/o EAM	3.45	3.80	3.33	3.67	3.56
w/o DAM	1.78	1.62	1.49	1.74	1.66
w/o DAP	3.84	4.20	4.10	4.38	4.13
$T_{\mathrm{d}} < T_{\mathrm{n}}$	1.52	1.31	1.37	1.28	1.23
EA-VAE	**4.02**	**4.54**	**4.69**	**4.65**	**4.48**

Table 3: The 5-scale MOS test ablation study results (↑) of "F02", "M07", "F04", and "M05" on the UASpeech for content similarity with mean scores and the 95% confidence intervals. The "w/o EAM", "w/o DAM", and "w/o DAP" refer to ablations where the Embedding Alignment Module, Distribution Alignment Module, and Duration Alignment Predictor are removed, respectively, and applying a time-stretching operation make $T_{\mathrm{d}} < T_{\mathrm{n}}$.

speech shows blurred formant structures and reduced high-frequency energy. Among the models, EA-VAE and Unit-DSR produce spectrograms that better preserve harmonic structure and energy distribution, especially for the speaker M07. In contrast, ASR-TTS, E2E-DSR, and ASA-DSR exhibit noticeable spectral degradation, with weaker or compressed features. These results suggest that EA-VAE achieves more accurate spectral reconstruction across different severity levels.

Comparsion with State-of-the-arts

To validate the effectiveness of the proposed EA-VAE model, we conduct both subjective and objective evaluations, comparing our approach with four representative state-of-the-art methods: two-stage and end-to-end models. The two-stage baseline includes ASR-TTS (Wang et al. 2024), while the end-to-end baselines comprise E2E-DSR (Wang et al. 2020), ASA-DSR (Wang et al. 2022), and UNIT-DSR (Wang et al. 2024).

As shown in Table 1, although some methods show progress in low and mid dysarthric speakers, the overall WER remains high, highlighting the challenges in DSR. Compared with existing DSR methods, EA-VAE achieved the lowest average WER of 62.19%, showing a 31.7% relative reduction from the original dysarthric speech. For low-intelligibility speakers (F02, M07), EA-VAE significantly reduced WER to 63.81% and 60.64%, outperforming all baselines. For mid-intelligibility speakers (F04, M05), it achieved 62.60% and 61.72%, maintaining consistent improvements.

As reported in Table 2, the original dysarthric speech has low intelligibility score (2.72), indicating that the speech quality is significantly impaired. Among baselines, Unit-DSR performs best score (4.39), followed by ASR-TTS (4.11) and E2E-DSR (3.83). EA-VAE achieves the highest average MOS of 4.48, outperforming all baselines. The result shows superior performance on M07, F04, and M05, and remains competitive on F02, producing relatively high-quality normal speech.

Ablation Study

As summarized in Table 3, without the DAM, performance drops significantly to 1.66, indicating its critical role in matching distributions. Removing the EAM, DAP, and setting $T_{\mathrm{d}} < T_{\mathrm{n}}$ reduces the score to 3.56, 4.13, and 1.23, respectively, demonstrating their significant contribution to modeling natural prosody. The results of the ablation experiments for MOS test are consistent with the results of the ablation experiments for WER test.

As presented in Table 4, removing EAM resulted in the largest increase in WER (94.46%), surpassing the original WER, indicating that the Distribution Alignment Mechanism contributes to improving speech clarity. Removing EMA, DAP and setting $T_{\mathrm{d}} < T_{\mathrm{n}}$ led to an improvement in the average WER, with values of 71.75%, 83.60%, and 63.43%, respectively, suggesting that these modules are essential for enhancing speech clarity. Meanwhile, the impact

Method	Low ↓		Mid ↓		Average / △ ↓
	F02 / △	M07 / △	F04 / △	M05 / △	
Original	95.9 / –	95.6 / –	81.7 / –	91.0 / –	91.05 / –
w/o EAM	71.45 / -25.5%	72.33 / -24.3%	70.21 / -14.1%	73.02 / -19.7%	71.75 / -21.2%
w/o DAM	92.60 / -3.5%	95.30 / +0.3%	93.85 / +14.9%	96.10 / +5.6%	94.46 / +3.7%
w/o DAP	**63.20 / -34.1%**	64.10 / -33.0%	62.85 / -23.1%	63.55 / -30.1%	63.43 / -30.3%
$T_{\text{d}} < T_{\text{n}}$	84.27 / -12.1%	79.53 / -16.8%	87.12 / +6.6%	83.47 / -8.3%	83.60 / -8.2%
EA-VAE	63.81 / -33.5%	**60.64 / -36.6%**	**62.60 / -23.4%**	**61.72 / -32.2%**	**62.19 / -31.7%**

Table 4: The ablation study results of WER on the UASpeech. Same notations as Table 1 and Table 3.

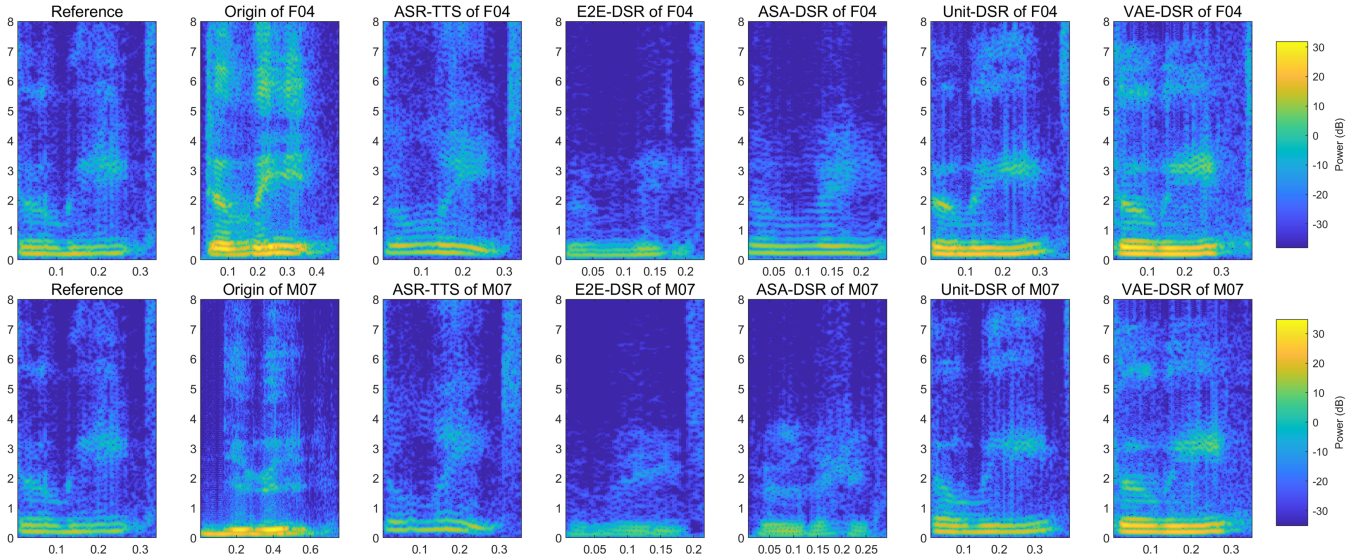

Figure 2: Spectrogram comparison of the five methods on the word "delete" for speakers F04 (Mid) and M07 (Low).

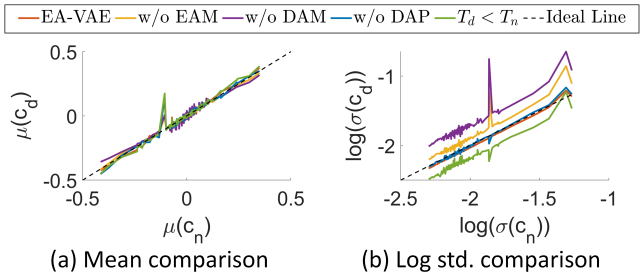

Figure 3: The comparison of means (a) and log standard deviation (b) of normal and dysarthric speech after removing different modules. The ideal line is represented as $\mu(c_{\text{d}}) = \mu(c_{\text{n}})$ and $log(\sigma(c_{\text{d}})) = log(\sigma(c_{\text{n}}))$ in (a) and (b). Same notations as Table 3.

of removing DAP on WER was relatively mild.

Figure 3(a) shows the mean comparison of 100 utterances generated by models with different modules removed. Removing each module does not result in significant changes in the mean, which is consistent with our expectations, indicating that the means of the same semantic content in dysarthric speech and normal speech is the same. Figure 3(b) presents the log standard deviation comparison of 100 utterances generated by models with different modules removed. When the DAM is removed, the log standard deviation exhibits noticeable fluctuations, suggesting a significant discrepancy between the distributions of normal speech and dysarthric speech. This phenomenon indicates that optimizing this distributional difference can enhance the quality of speech synthesis. When $T_{\text{d}} < T_{\text{n}}$, the log standard deviation is represented as $log(\sigma(c_{\text{n}})) > log(\sigma(c_{\text{d}}))$, further confirming the necessity of ensuring $T_{\text{d}} \geq T_{\text{n}}$ during model alignment.

Conclusion

This paper introduces EA-VAE, a simple yet effective DSR model that only leverages a VAE to align the frame-level embeddings between dysarthric and normal speech for reconstruction. EA-VAE reduces the significant acoustic pattern differences between dysarthric and normal speech by aligning frame-level embeddings, minimizing prior distributional differences, and modeling duration patterns. Rigorous experiments on UASpeech corpus demonstrate the superiority of our proposed method, with a 31.7% relative WER reduction and the highest subjective MOS score (4.48). In addition, ablation experiments demonstrate the effectiveness of the three alignment mudules. In future work, we aim to enhance the generalization ability of the model across different dysarthric severity levels and speaker identities.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFB2603902, in part by the Major Science and Technology Specific Project of Xining under Grant 2024-Z-7.

References

Aihara, R.; Takashima, R.; Takiguchi, T.; and Ariki, Y. 2012. Consonant enhancement for articulation disorders based on non-negative matrix factorization. In *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 1–4. IEEE.

Aihara, R.; Takashima, R.; Takiguchi, T.; and Ariki, Y. 2013. Individuality-preserving voice conversion for articulation disorders based on non-negative matrix factorization. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 8037–8040. IEEE.

Aihara, R.; Takiguchi, T.; and Ariki, Y. 2017. Phoneme-Discriminative Features for Dysarthric Speech Conversion. In *Proc. INTERSPEECH*, 3374–3378.

Almadhor, A.; Irfan, R.; Gao, J.; Saleem, N.; Rauf, H. T.; and Kadry, S. 2023. E2E-DASR: End-to-end deep learning-based dysarthric automatic speech recognition. *Expert. Syst. Appl.*, 222: 119797.

Biadsy, F.; Weiss, R. J.; Moreno, P. J.; Kanevsky, D.; and Jia, Y. 2019. Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation. *arXiv preprint arXiv:1904.04169*.

Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6): 1505–1518.

Chen, X.; Wang, Y.; Wu, X.; Wang, D.; Wu, Z.; Liu, X.; and Meng, H. 2024a. Exploiting audio-visual features with pretrained av-hubert for multi-modal dysarthric speech reconstruction. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 12341–12345. IEEE.

Chen, Y.; Yue, X.; Gao, X.; Zhang, C.; D'Haro, L. F.; Tan, R. T.; and Li, H. 2024b. Beyond Single-Audio: Advancing Multi-Audio Processing in Audio Large Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 10917–10930. Miami, Florida, USA: Association for Computational Linguistics.

Chen, Y.; Yue, X.; Zhang, C.; Gao, X.; Tan, R. T.; and Li, H. 2024c. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*.

Chen, Z.; Ramabhadran, B.; Biadsy, F.; Zhang, X.; Chen, Y.; Jiang, L.; Chu, F.; Doshi, R.; and Moreno, P. J. 2021. Conformer Parrotron: A Faster and Stronger End-to-End Speech Conversion and Recognition Model for Atypical Speech. In *Proc. INTERSPEECH*, 4828–4832.

Chu, M.; Yang, M.; Xu, C.; Ma, Y.; Wang, J.; Fan, Z.; Tao, Z.; and Wu, D. 2023. E-DGAN: An encoder-decoder generative adversarial network based method for pathological to normal voice conversion. *IEEE J. Biomed. Health Inform.*, 27(5): 2489–2500.

Doshi, R.; Chen, Y.; Jiang, L.; Zhang, X.; Biadsy, F.; Ramabhadran, B.; Chu, F.; Rosenberg, A.; and Moreno, P. J. 2021. Extending parrotron: An end-to-end, speech conversion and speech recognition model for atypical speech. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 6988–6992. IEEE.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Imai, S.; Nose, T.; Kanagaki, A.; Watanabe, S.; and Ito, A. 2020. Improving pronunciation clarity of dysarthric speech using cyclegan with multiple speakers. In *Proc. IEEE 9th Global Conf. Consum. Electron.*, 366–367. IEEE.

Kain, A. B.; Hosom, J.-P.; Niu, X.; Van Santen, J. P.; Fried-Oken, M.; and Staehely, J. 2007. Improving the intelligibility of dysarthric speech. *Speech Commun.*, 49(9): 743–759.

Kim, H.; Hasegawa-Johnson, M.; Perlman, A.; Gunderson, J. R.; Huang, T. S.; Watkin, K. L.; and Frame, S. 2008. Dysarthric speech database for universal access research. In *Proc. INTERSPEECH*, volume 2008, 1741–1744.

Kim, J.; Kim, S.; Kong, J.; and Yoon, S. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Adv. Neural. Inf. Process. Syst.*, 33: 8067–8077.

Kim, J.; Kong, J.; and Son, J. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proc. Int. Conf. Mach. Learn.*, 5530–5540. PMLR.

Kong, J.; Kim, J.; and Bae, J. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Adv. Neural. Inf. Process. Syst.*, 33: 17022–17033.

Kumar, S. A.; and Kumar, C. S. 2016. Improving the intelligibility of dysarthric speech towards enhancing the effectiveness of speech therapy. In *Proc. Int. Conf. Adv. Comput. Commun. Informat.*, 1000–1005. IEEE.

Lee, J.; Littlejohn, M. A.; and Simmons, Z. 2017. Acoustic and tongue kinematic vowel space in speakers with and without dysarthria. *Int. J. Speech-Language Pathol.*, 19(2): 195–204.

Li, J.; Lavrukhin, V.; Ginsburg, B.; Leary, R.; Kuchaiev, O.; Cohen, J. M.; Nguyen, H.; and Gadde, R. T. 2019. Jasper: An end-to-end convolutional neural acoustic model. *arXiv preprint arXiv:1904.03288*.

Liu, D.; Lin, Y.; Bu, H.; and Li, M. 2024. Two-stage and self-supervised voice conversion for zero-shot dysarthric speech reconstruction. In *Proc. Int. Conf. Asian Lang. Process.*, 423–427. IEEE.

Matsubara, K.; Okamoto, T.; Takashima, R.; Takiguchi, T.; Toda, T.; Shiga, Y.; and Kawai, H. 2021. High-intelligibility speech synthesis for dysarthric speakers with LPCNet-based TTS and CycleVAE-based VC. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 7058–7062. IEEE.

Mehrez, H.; Chaiani, M.; and Selouani, S. A. 2024. Using StarGANv2 voice conversion to enhance the quality of dysarthric speech. In *Proc. Int. Conf. Artif. Intell. Inf. Commun.*, 738–744. IEEE.

Prananta, L.; Halpern, B. M.; Feng, S.; and Scharenborg, O. 2022. The effectiveness of time stretching for enhancing dysarthric speech for improved dysarthric speech recognition. In *Proc. INTERSPEECH*.

Prenger, R.; Valle, R.; and Catanzaro, B. 2019. Waveglow: A flow-based generative network for speech synthesis. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 3617–3621. IEEE.

Purohit, M.; Patel, M.; Malaviya, H.; Patil, A.; Parmar, M.; Shah, N.; Doshi, S.; and Patil, H. A. 2020. Intelligibility improvement of dysarthric speech using mmse discogan. In *Proc. Int. Conf. Signal Process. Commun.*, 1–5. IEEE.

Rudzicz, F. 2011. Acoustic transformations to improve the intelligibility of dysarthric speech. In *Proc. Second Workshop Speech Lang. Process. Assist. Technol.*, 11–21.

Shahamiri, S. R.; Lal, V.; and Shah, D. 2023. Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system. *IEEE Trans. Neur. Sys. Reh.*

Shen, J.; Pang, R.; Weiss, R. J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 4779–4783. IEEE.

Ueno, S.; Lee, A.; and Kawahara, T. 2024. Refining Synthesized Speech Using Speaker Information and Phone Masking for Data Augmentation of Speech Recognition. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Adv. Neural. Inf. Process. Syst.*, 30.

Wang, D.; Liu, S.; Wu, X.; Lu, H.; Sun, L.; Liu, X.; and Meng, H. 2022. Speaker identity preservation in dysarthric speech reconstruction by adversarial speaker adaptation. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 6677–6681. IEEE.

Wang, D.; Yu, J.; Wu, X.; Liu, S.; Sun, L.; Liu, X.; and Meng, H. 2020. End-to-end voice conversion via cross-modal knowledge distillation for dysarthric speech reconstruction. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 7744–7748. IEEE.

Wang, Y.; Wu, X.; Wang, D.; Meng, L.; and Meng, H. 2024. Unit-dsr: Dysarthric speech reconstruction system using speech unit normalization. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 12306–12310. IEEE.

Yamagishi; Junichi, V.; Christophe, M.; and Kirsten. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92). *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*.

Yang, S. H.; and Chung, M. 2020. Improving dysarthric speech intelligibility using cycle-consistent adversarial training. *arXiv preprint arXiv:2001.04260*.

Yu, C.; Su, X.; and Qian, Z. 2023. Multi-stage audio-visual fusion for dysarthric speech recognition with pre-trained models. *IEEE Trans. Neur. Sys. Reh.*, 31: 1912–1921.

Zhang, D.; Zhang, H.; Lu, W.; Li, W.; Wang, J.; and Wei, J. 2025. Long-range and Non-stationary Encoding for Dysarthric Speech Data Augmentation. *IEEE J. Sel. Top. Signal Process.*, 1–16.

Zheng, W.-Z.; Han, J.-Y.; Chen, C.-Y.; Chang, Y.-J.; and Lai, Y.-H. 2023. Improving the efficiency of dysarthria voice conversion system based on data augmentation. *IEEE Trans. Neur. Sys. Reh.*, 31: 4613–4623.