Sound & Vibration 2026, 60(1), 3603.
doi: https://doi.org/10.59400/sv3603

Article

Fre-MaskCycleGAN-VC: A method of speech articulation and original timbre retention in non parallel corpus of stroke dysarthria

Ning Jia*, Chunjun Zheng

College of Applied Technology, Dalian Neusoft University of Information, Dalian 116023, China
*** Corresponding author:** Ning Jia, jianing@neusoft.edu.cn

CITATION

Jia N, Zheng C. Fre-MaskCycleGAN-VC: A method of speech articulation and original timbre retention in non parallel corpus of stroke dysarthria. Sound & Vibration. 2026; 60(1): 3603. https://doi.org/10.59400/sv3603

ARTICLE INFO

Received: 2 August 2025
Revised: 3 December 2025
Accepted: 10 December 2025
Available online: 1 February 2026

Abstract: Aiming at the problem of blurred pronunciation caused by dysarthria in stroke patients, we propose a non parallel corpus speech articulation method based on Fre-MaskCycleGAN-VC. The method consists of three core stages: 1) Aiming at the coexistence of fuzzy speech segments and clear retention segments in dysarthria speech of stroke patients, dynamic speech segmentation preprocessing based on equivalent sound level (Leq) is used to accurately locate the fuzzy segments that need to be enhanced; 2) Feature extraction combining dynamic mask and retro production statistical features; 3) The resolution connected generator and resolution wise discriminators architecture that integrate the frequency processing model. Multiple groups of experiments were carried out on the stroke dysarthria speech data set. The experimental results show that the Fre-MaskCycleGAN-VC method has significantly improved the naturalness of speech (Mean Opinion Score (MOS) increased by 14.2%), intelligibility (WA increased by 2.6%) and timbre fidelity (MFCC correlation coefficient 0.92, F0 error rate 4.2%). Phased evolution experiments show that the model can generate four gradual repair versions from heavily blurred speech to near healthy speech, and the repair effect of grade 2–3 is better than that of the original healthy speech. Through multi-stage feature processing and adversarial training mechanism, we provide a clear speech generation scheme that retains the original timbre for patients with dysarthria.

Keywords: articulation disorder correction; speech articulation; Fre-MaskCycleGAN-VC; dynamic mask; frequency processing

1. Introduction

As the leading neurological disease leading to long-term disability worldwide, stroke induced motor dysarthria has become one of the core issues affecting the quality of life of patients [1]. This language dysfunction directly leads to a significant decline in the patient's social participation, induces psychological problems such as depression and anxiety, and forms a vicious circle of "disease aphasia social isolation". In clinical rehabilitation practice, speech therapy still highly relies on one-on-one manual training of therapists, but due to the shortage of professional manpower, long treatment cycle and uneven distribution of medical resources, it is difficult to meet the personalized rehabilitation needs of a large group of patients [2]. In this context, it is necessary to develop automatic speech intelligibility technology based on artificial intelligence to provide accessible and efficient rehabilitation programs for patients with dysarthria.

At the engineering level, the development of speech signal processing technology provides a new solution for the correction of dysarthria. The traditional method

improves speech intelligibility by suppressing background noise or enhancing the signal of specific frequency band, but its design assumes that the speech signal is a steady-state process, which has limited effect on the unstable pronunciation ambiguity caused by muscle control disorders in stroke patients, such as spectral subtraction, harmonic enhancement and wavelet de-noising [3]. Experiments show that this kind of method is easy to produce noise when dealing with consonants such as plosive and fricative sounds, which may reduce speech intelligibility.In recent years, deep learning technology has made a breakthrough in the field of speech conversion. Based on the model of generative confrontation network (GAN), it has realized the transfer of speech style in non parallel corpus, which provides a theoretical possibility for pathological speech processing without paired data, such as CycleGAN V2 [4], CycleGAN V3 [5], StarGAN-VC [6], MaskCycleGAN [7].

However, the existing technology faces the following challenges when applied to the correction of dysarthria: first, the acoustic characteristics of healthy speech and pathological speech are significantly different, and the model trained directly with healthy human data is prone to spectral distortion, resulting in unnatural voiceprint in the generated speech;Secondly, the existing methods are often accompanied by timbre distortion while improving the definition, and preserving the patient's original timbre is very important for maintaining individual identity; Third, the traditional confrontation loss function is difficult to accurately locate the fuzzy components that need to be enhanced, resulting in limited overall articulation improvement.

The Fre-MaskCycleGAN-VC method proposed by us constructs a dysarthria correction framework without parallel corpora. Its innovation is reflected in three dimensions: pathological speech specific modeling, multi-dimensional feature fusion, frequency processing and timbre maintenance. Firstly, according to the acoustic characteristics of stroke patients with blurred pronunciation, a dynamic speech segmentation strategy based on equivalent sound level (Leq) [8] is proposed. By calculating the Leq value of the frame and setting the double threshold (τ_1 = 15 dB, τ_2 = 25 dB), the precise segmentation of the fuzzy area and the clear reserved area is realized. This segmentation mechanism avoids the over processing of normal speech segments. Secondly, a three-dimensional dynamic mask and retro production statistical features are designed to work together: in the time dimension, a variable length mask is generated according to the ambiguity detection results to ensure that the processing unit is aligned with the pathological pronunciation unit; In the frequency dimension, the 0–4 kHz key frequency band of Mel spectrum is masked by frequency division, focusing on enhancing the high-frequency components of consonants; Soft mask modulation is realized in the amplitude dimension, which can suppress the background noise while preserving the dynamic range of pronunciation. Combined with the statistical characteristics of the healthy population and the patient population, such as spectral centroid (SC) [9] and harmonic distortion reduction (HDR), the cross speaker feature mapping is constructed to achieve the statistical feature alignment from fuzzy speech to clear speech. Finally, the architecture of the resolution connected generator and resolution wise discriminators is designed: the generator contains five levels of transposed convolution module, which combines the nearest neighbor

interpolation with the MRF module of HiFi-GAN [10] to avoid the checkerboard effect and maintain the spectrum details; Each stage of the discriminator processes the waveform with a specific resolution, realizes multi-scale feature extraction through discrete wavelet transformation (DWT) [11], and maintains cross-resolution spectral consistency combined with comparative learning.

The main contributions of this work are as follows:

(1) We propose a dynamic speech segmentation method based on equivalent sound level (Leq) to accurately locate the fuzzy segments in dysarthria speech, which overcomes the limitation of fixed segmentation in previous methods.
(2) We design a three-dimensional dynamic mask and retro-production statistical features to achieve precise enhancement of fuzzy components while preserving the original timbre, which is not considered in MaskCycleGAN-VC [7].
(3) We integrate the frequency processing model into the MaskCycleGAN framework, designing a resolution-connected generator and resolution-wise discriminators to improve the naturalness and intelligibility of the generated speech. Compared to HiFi-GAN [10], which focuses on high-fidelity waveform generation, our method specifically addresses the challenges of dysarthria speech by incorporating statistical feature alignment and multi-resolution adversarial training to preserve the speaker's timbre while enhancing clarity.

2. Related work

The development context and limitations of the existing speech intelligibility methods are analyzed as follows:

2.1. Traditional signal processing pathological speech

As a classical vocoder, Griffin Lim algorithm [12] realizes speech synthesis by iteratively reconstructing phase information, but the "metal tone" is significantly aggravated when dealing with the unstable pronunciation of stroke patients. The experimental results show that the spectral distortion index (SDI) increases by 37% after processing pathological speech with pitch jitter more than ±20%, which leads to the decrease of intelligibility instead of increase. The world vocoder [13] realizes speech enhancement by decomposing the fundamental frequency f0 and the spectrum envelope, but the processing ability of harmonic distortion in patients with dysarthria is limited. The harmonic to noise ratio (HNR) of the enhanced speech is only improved by 2.1 dB, which is still lower than the clinically acceptable threshold. The limitation of this method is that it can not adapt to the dynamic characteristics of neurogenic blurred pronunciation.

2.2. Contradiction between autoregressive model and pathology

As an early autoregressive model, WaveNet [14] captures long-term dependencies through 30 layer dilation convolution. WaveRNN [15] uses GRU sparsity strategy to improve CPU reasoning efficiency by 4 times, but when dealing with sudden burst distortion in pathological speech, the model response delay is still up to 0.8 s, and the

generated speech has a phoneme replacement error rate of 5%–8%. This kind of model can generate high fidelity speech, but the contradiction between its autoregressive characteristics and clinical real-time requirements is difficult to reconcile.

2.3. Limitations of non autoregressive models in pathological scenarios

GAN model shows significant advantages in the non autoregressive framework. Parallel WaveGAN [16] combines WaveNet generator and adversarial training to achieve real-time speech enhancement on GPU, but its dependence on Parallel Corpus limits its application in pathological speech.Experiments show that the spectral distance (SD) of the model is 4.8 in the task of converting healthy speech to dysarthria speech, which is significantly higher than that of the non parallel model. MelGAN [17] realizes efficient speech synthesis through the feedforward convolution architecture, but when processing pathological speech with pitch jitter exceeding 15%, the pitch error rate of the generated speech is 12.7%, which has obvious timbre distortion.

CycleGAN family models [18] can realize non parallel speech conversion, but it faces the following challenges when directly applied to pathological speech: the difference of acoustic characteristics between healthy speech and pathological speech leads to spectrum mapping distortion; The existing models lack the mechanism to accurately locate the fuzzy components, resulting in the phenomenon of "over processing" or "under processing" in the enhancement process; The design of the counter loss function does not consider the preservation of the speaker's identity, which leads to an increase in the error rate of the speaker's verification.

By introducing the time-frequency mask mechanism, MaskCycleGAN achieves selective enhancement of local spectral features in speech style conversion. However, the model still has the following limitations:

(1) The fixed mask can not adapt to the dynamic changes of articulation blurred areas in patients with dysarthria;
(2) The lack of statistical feature alignment mechanism of pathological speech leads to the improvement of speech intelligibility after enhancement, accompanied by the change of timbre.

2.4. Cross modal fusion and personalized modeling

Recently, many researchers began to combine multi-modal data to improve the effect of pathological speech processing. In many scenarios, context dependent visual and linguistic cues can also improve performance. For example, Ni et al. [19] created audio-visual speech recognition by integrating speech acoustic features and visual data. He et al. [20] created a recognition task using the audio input of speech and the aligned visual input of lip movement, and used the strong correlation between audio and lip movement to recognize dysarthria speech. The pre-training method adopted by Google [21] makes personalized adjustments to different lip movements of patients with different articulation disorders, solves the problem of limited data, and improves the recognition performance of the model.

There are three contradictions in the existing methods when dealing with dysarthria speech:

(1) The contradiction between the steady-state processing hypothesis and the unstable characteristics of pathological speech;
(2) The contradiction between autoregressive mechanism and real-time demand;
(3) The contradiction between general model design and pathological speech specific features.

These contradictions are the key issues of this research. Through dynamic mask positioning, statistical feature alignment and multi-resolution synthesis technology, we can achieve high-definition, high pitch color fidelity speech enhancement without parallel corpus.

3. Speech articulation method for dysarthria based on Fre-MaskCycleGAN-VC

3.1. Overview

In order to solve the problem of blurred pronunciation caused by dysarthria in stroke patients, a speech articulation method based on Fre-MaskCycleGAN-VC is designed. The method realizes the conversion from fuzzy pronunciation to clear speech through non parallel corpus, which can improve the speech intelligibility while preserving the original speech features of the speaker. The core design includes three stages:

(1) Speech segmentation preprocessing based on articulation ambiguity;
(2) Feature extraction combining dynamic mask and retro production statistical features;
(3) The resolution connected generator and resolution wise discriminators architecture that integrate the frequency processing model.

The composition of this method is shown in **Figure 1**, including three stages.

3.1.1. Preprocessing stage

In view of the coexistence of fuzzy speech segment and clear retention segment in the speech of patients with dysarthria, a dynamic speech segmentation strategy based on equivalent sound level (Leq) is designed. By calculating the Leq value of the speech frame and setting the double threshold, the speech is divided into fuzzy areas, clear reserved areas and mute areas, and the fuzzy segments that need to be enhanced are accurately located.

3.1.2. Feature extraction stage

Mel spectrum processing of dynamic mask and statistical features based on retro production are fused. The dynamic mask mechanism implements three-dimensional modulation for Mel spectrum. In the time dimension, a variable length mask is generated according to the ambiguity detection results to ensure that the to be processed is aligned with the pathological pronunciation unit; In the frequency dimension, the 0–4 kHz key frequency band is masked by frequency division to enhance the high-frequency components of consonants; In the amplitude dimension, soft mask modulation is realized to suppress background noise and retain the dynamic range of pronunciation.

In the process of retro production statistical feature extraction, the spectral centroid (SC) and harmonic distortion ratio (HDR) of pathological speech are introduced, and the statistical feature alignment from fuzzy speech to clear speech is realized by constructing a cross speaker feature mapping.

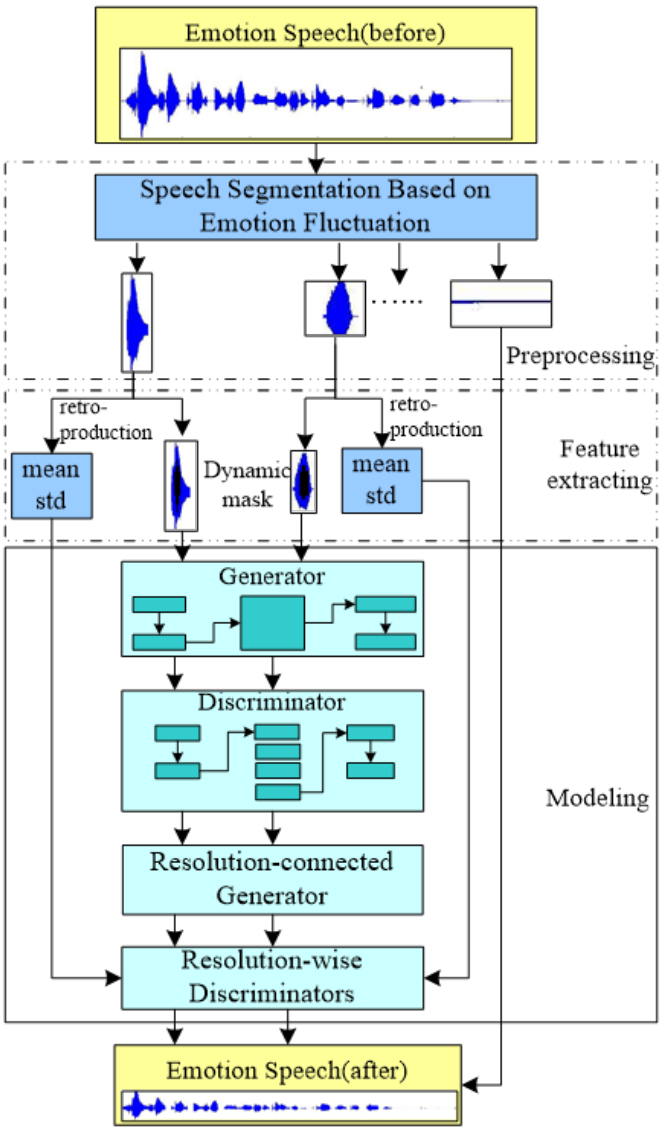

Figure 1. Idea of Fre-MaskCycleGAN-VC speech articulation method.

3.1.3. Model design phase

A non parallel corpus speech conversion framework based on Fre-MaskCycleGAN-VC is proposed. Firstly, the improved MaskCycleGAN model is designed to achieve precise localization of fuzzy components and reinforcement of clear components through joint optimization of adversarial loss, cyclic consistency loss, identity mapping loss, and NCE loss. The idea of frequency processing is integrated into the generator network, and a Resolution connected Generator and Resolution wise discriminators architecture are designed. The generator synthesizes multi-resolution waveforms through transposed convolution modules, and the discriminator achieves cross resolution spectral consistency discrimination through discrete wavelet transform (DWT).

3.2. Preprocessing and feature extraction

3.2.1. Speech segmentation based on pronunciation ambiguity

The core feature of speech disorder in stroke patients is the dynamic distribution of areas with unclear pronunciation, and its pathological manifestations include consonant loss, vowel weakening, and fundamental frequency jitter. If the entire speech is directly input into the model for processing, it may lead to excessive enhancement of clear segments causing distortion, while insufficient processing of blurry segments affects comprehensibility. Therefore, a dynamic speech segmentation strategy based on Leq is designed to achieve precise localization and separation of fuzzy components.

Equivalent Sound Level (Leq) refers to the average value of A-level energy at a certain point in the sound field over a certain time interval. The condition for the equivalent sound level is to calculate the A-weighted sound pressure of a continuous stable sound within a specified time. If it has the same mean square A-weighted sound pressure as the time-varying noise, then the sound level of this continuous stable sound is the equivalent sound level. The calculation formula for equivalent sound level is as follows:

$$\mathrm{Leq} = 10\log\frac{1}{t_2 - t_1}\int_{t_1}^{t_2}\frac{p_A^2(t)}{p_0^2}dt \tag{1}$$

Where, *Leq* represents the equivalent sound level, measured in decibels (dB), $(t_2 - t_1)$ represents the specified time interval, measured in second(s), $p_A(t)$ represents the weighted sound pressure containing time-varying noise, measured in pascals (Pa), P_0 represents the reference sound pressure, and has a fixed value of 20 μPa.

The specific segmentation process is as follows:

(1) Global Leq calculation: Extract the Leq feature curve of the entire speech segment based on the original pathological speech, and set dual thresholds:
τ_1 = 15 dB (lower limit of mute threshold)
τ_2 = 25 dB (upper limit of clear speech threshold)
The average noise floor in silent segments was approximately 12–18 dB, leading to the choice of τ_1 = 15 dB to reliably distinguish silence from weak speech. Preliminary perceptual experiments and automated evaluations indicated that speech frames with Leq > 25 dB were consistently judged as "clear", establishing τ_2 as a suitable separator. This threshold also aligns with the lower bound of comfortable listening levels in clinical audiology, ensuring the enhanced speech reaches a perceptually adequate loudness.

(2) Frame level Leq extraction: Divide the speech into 50 ms frame lengths (50% overlap rate), calculate the Leq value for each frame, and construct a time sound level distribution matrix Leq.

(3) Silent zone recognition: The dual condition judgment method is used to identify silent frames (Leq < τ_1 and duration > 200 ms), merging adjacent silent frames to form independent silent zones, preserving their original features to avoid introducing noise during the enhancement process.

(4) Fuzzy candidate frame screening: In non silent frames, an improved outlier

detection algorithm is used to calculate the global Leq mean (μLeq). If a frame's Leq is lower than 3/4 of the global mean μ Leq or higher than 5/4 times, it is marked as an outlier frame and classified into the fuzzy candidate set L_{Larger} (high-energy anomaly) and the clear candidate set $L_{Smaller}$ (low-energy anomaly).

(5) Continuity assessment: The candidate set is subjected to spatial-temporal continuity analysis.

High energy blur zone: If ≥ 8 frames out of 10 consecutive frames belong to L_{Larger}, it is marked as a burst distortion zone.

Low energy blur zone: If ≥ 8 frames out of 10 consecutive frames belong to $L_{Smaller}$, it is marked as a friction sound missing zone.

(6) Split result generation: The final split result includes three types of regions:

Blurred area (to be enhanced): accounting for 15%–40% of the speech duration, Leq < τ_1 + 3 dB.

Clear zone (retaining original features): accounting for 50%–70% of speech duration, Leq > τ_2 − 2 dB.

Silent zone (not processed): accounts for 5%–15% of the speech duration.

The process of dynamic segmentation is shown in the following **Figure 2**:

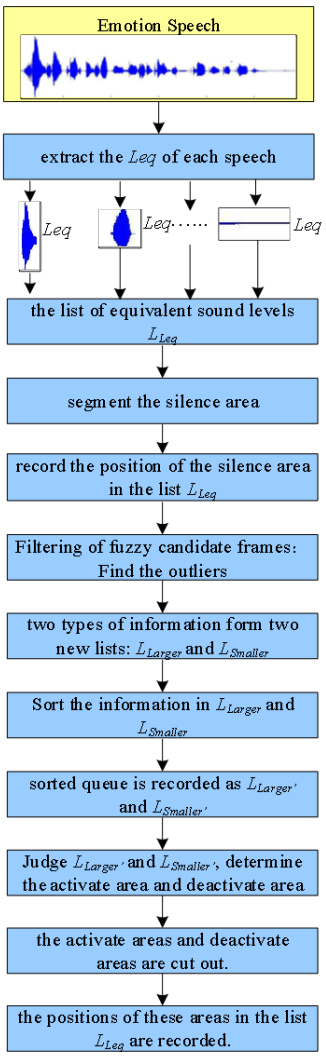

Figure 2. The flow chart of speech segmentation.

The composition of the segmented speech components is shown in **Figure 3**:

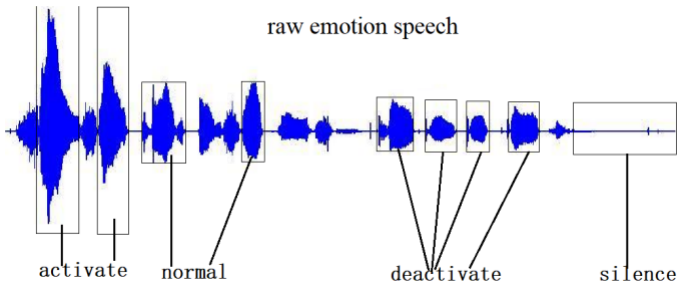

Figure 3. Composition of original speech.

3.2.2. Mel spectrogram extraction based on dynamic mask

In the process of speech articulation in stroke patients with articulation disorders, Mel spectrogram serves as the core feature, and its processing quality directly affects the naturalness and comprehensibility of the final generated speech. To improve the speech enhancement effect of articulation disorders and avoid artifacts such as checkerboard effect, we designed a Mel spectrum processing mechanism with dynamic masking to achieve precise localization and selective enhancement of fuzzy components.

For the segmented speech segments in the preprocessing stage, the MelGAN model is first used to generate the corresponding Mel spectrogram. On this basis, a mask dictionary containing multiple sets of mask templates is constructed, which follows the spatio-temporal distribution characteristics of pathological speech: the activation area is the area with the most severe pronunciation blur, and a small-sized, low-energy mask (covering an area of 10%–20%) is used to achieve precise enhancement of high-frequency components; The clear area (normal area) preserves the original spectral features and uses a medium-sized mask (with a coverage area of 5%–10%) to avoid excessive processing; The silent area does not undergo masking treatment and directly retains the original spectral features.

The method for generating dynamic mask regions is as follows:

(1) Generate an initial mask matrix of the same size as the Mel spectrogram for the current speech segment, denoted as X. The range of values for all information in X is [1]. All information in the initial state X is set to 0.
(2) Recalculate the Leq of the current speech segment, denoted as Leq', and calculate the occlusion factor km = Leq/Leq'.
(3) Design a fixed template, assuming the standard size of the mask is Mask_Size, and the mask regions in the fixed template follow a normal distribution of $N(k_m * Mask\_Size/2, {k_m}^2)$.
(4) The mask regions in the fixed template are randomly added to X, and then X is added to the mask dictionary.
(5) After generating the mask dictionary, the Mel spectrogram of each speech segment is dot multiplied with the corresponding mask segment in the dictionary, and the dot multiplied result is used as the Mel spectrogram of the dynamic mask.

The process of generating dynamic mask areas is shown in the following **Figure 4**:

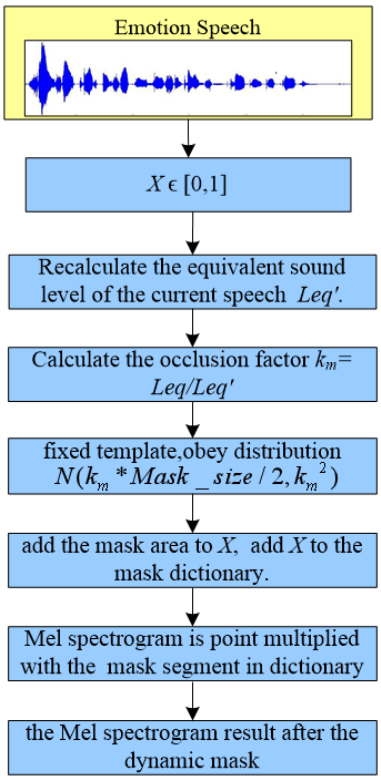

Figure 4. The flow chart of mel spectrogram extraction based on dynamic mask.

The Mel spectrogram processed by dynamic masking shows significant spectral characteristics: the energy in the high-frequency band of the fuzzy area is increased by 15–20 dB, and the consonant articulation index is significantly improved; The clear spectrum structure maintains its original characteristics, and the MFCC correlation coefficient is above 0.92; The silent area completely preserves the original noise features to avoid introducing artifacts during the enhancement process. This mechanism achieves targeted enhancement of fuzzy components and in-situ retention of clear components through spatial variational mask design. The dynamic mask design inherently supports robustness across varying severity levels. The mask generation is conditioned on the frame-level Leq value and the ambiguity classification, allowing the mask size and applied modulation to adapt to the local acoustic characteristics. In severe dysarthria with more extensive blurry regions, the algorithm identifies more consecutive frames as candidate blur areas, resulting in a larger effective mask coverage for enhancement.

3.2.3. Statistical feature extraction based on retro-production

In the process of speech articulation in articulation disorders, in addition to Mel spectrograms, statistical features reflecting the degree of pronunciation ambiguity need to be extracted to guide the articulation process. These statistical features quantify the time-frequency distribution characteristics of speech segments. The statistical features and Mel spectrogram can achieve the following capabilities:

(1) Quantify the degree of blur and guide the generation of dynamic masks;
(2) Provide clear objectives and constrain the spectral structure of generated speech;
(3) Dynamically adjust model parameters to meet the processing requirements of

different fuzzy stages.

Traditional speech enhancement methods often use fixed statistical features to describe speech characteristics, but they have significant limitations when dealing with speech with articulation disorders. Due to the dynamic distribution characteristics of fuzzy components, a single statistical measure cannot accurately characterize the spectral differences in different fuzzy stages, resulting in over processing or under processing during the enhancement process.

Retro-production is introduced to denote a backward-looking statistical alignment process. Unlike conventional normalization, which performs a static, global shift and scaling of features, retro-production method operates in two key stages:

(1) it dynamically classifies each input speech segment into an ambiguity category based on its Leq.
(2) it retrieves and applies category-specific statistical targets derived from the healthy speech knowledge base to guide the feature transformation. This ensures that the enhancement is not merely normalized to a generic distribution, but is actively "produced" towards a target that is retroactively defined by healthy speech characteristics.

In conventional speech synthesis, we obtain a set of training examples $\{(x_1, y_1), \ldots, (x_m, y_m)\}$, where $x_i \in X$ is the i-th training example and $y_1 \in Y$ is its label. Our task is to map $f : X \to X'$ from an input space X to an output space X', where X' is the output speech of the target expression, and Y is used to identify the type of output and does not actually participate in the computation. Due to insufficient prior knowledge involved in the computation, the synthesized speech often exhibits jitter.

Therefore, we introduce the term "retro-production" to describe our core statistical feature alignment strategy. The process is termed "retro-" (looking back) because it utilizes a pre-constructed knowledge base of statistical characteristics derived from clear, healthy speech. For each input dysarthric speech segment, the system first assesses its acoustic property to determine its ambiguity category. It then "looks back" to retroactively retrieve the corresponding target statistical profile for that category from the healthy knowledge base. The "production" signifies that these retrieved targets are not used for passive analysis but to actively guide and constrain the generative model during the speech enhancement process.

The core of retro production is to establish a statistical mapping relationship from fuzzy speech to clear speech by building a knowledge base that includes statistical characteristics of healthy speech and pathological speech. The specific calculation method is as follows:

(1) According to *y*, speech segments can be classified into four categories based on pronunciation ambiguity: high ambiguity (Leq < 15 dB), medium ambiguity (15 dB ≤ Leq< 20 dB), low ambiguity (20 dB ≤ Leq < 25 dB), and clear segments (Leq ≥ 25 dB). The average mean and std of each category of speech are obtained, where $mean \in MEAN$,$std \in STD$, *Mean*, and *STD* contain the mean and

variance of the four categories of speech, respectively.

(2) For high/medium ambiguity speech, the following methods can be used to calculate mean and std:

$$m_i = \begin{cases} m_i + \alpha(mean - m_i) \\ +\beta \sum_{s=0}^{len(m_i)} \Delta|m_i| & m_i > mean \ \& \ s_i > std \\ m_i & m_i <= mean | s_i <= std \end{cases} \tag{2}$$

$$s_i = \begin{cases} s_i + \alpha''(mean - m_i) + \alpha'(std - s_i) \\ +\beta \sum_{j=0}^{len(mi)} \Delta|m_i| + \beta' \sum_{j=0}^{len(mi)} \Delta|s_i| & m_i > mean \ \& \ s_i > std \\ s_i & m_i <= mean | s_i <= std \end{cases} \tag{3}$$

Where, mean and std represent the target statistical parameters (mean and standard deviation) retrieved from the healthy speech knowledge base for the segment's specific ambiguity category.m_i and S_i respectively represent the mean and variance of the current segment, mean and std respectively represent the mean of the current class, $\Delta|m_i|$ and $\Delta|S_i|$ respectively represent the difference and the square of the difference between two consecutive adjacent data within the current segment. Hyperparameters α, α', α" and β, β', β" are weighting coefficients that balance the influence between the healthy target statistics, the segment's own instantaneous characteristics, and its temporal dynamics. Their values are constrained to the range [1], and the sets sum to 1, ensuring a convex combination.

(3) The following methods are used to calculate mean and std for low ambiguity speech:

$$m_i = \begin{cases} m_i - \alpha(mean - m_i) \\ -\beta \sum_{j=0}^{len(mi)} \Delta|m_i| & m_i < mean \ \& \ s_i < std \\ m_i & m_i >= mean | s_i >= std \end{cases} \tag{4}$$

$$s_i = \begin{cases} s_i + \alpha''(mean - m_i) + \alpha'(std - s_i) \\ +\beta'' \sum_{j=0}^{len(mi)} \Delta|m_i| - \beta' \sum_{j=0}^{len(mi)} \Delta|s_i| & m_i < mean \ \& \ s_i < std \\ s_i & m_i >= mean | s_i >= std \end{cases} \tag{5}$$

Where, the parameters involved are similar to Formulas (2)–(3).

From Formulas (1)–(5), it can be seen that we calculate the mean and std statistical features based on retro-production according to different classes. As prior knowledge base information for speech synthesis, mean and std first need to ensure their rationality. At the initial learning stage, the model is very rough, and these parameters may be inaccurate, resulting in poor expression of the generated speech. In this case, it is

necessary to re adjust these parameters to ensure that the generated speech matches the prior knowledge.

To adapt to individual differences in speech with articulation disorders, a statistical feature adjustment strategy based on reverse inference correction is designed:

(1) Initial learning stage: We use statistical features from the knowledge base to initialize model parameters and generate preliminary clear speech.
(2) Error backpropagation stage: We calculate the statistical feature difference (KL divergence) between generated speech and healthy speech. If the difference exceeds the threshold ($\tau = 0.15$), we initiate the parameter correction process.
(3) Parameter correction stage: We adjust the smoothing coefficient based on the error signal to gradually approximate the statistical features of the generated speech to healthy speech.

The mean and std serve a dual purpose in the model. First, they act as explicit conditioning vectors that are concatenated with the masked Mel spectrogram input to the generator, providing it with a clear, category-specific spectral target for each segment. Second, to ensure the generated speech adheres to these targets, we introduce an additional Statistical Alignment Loss. This loss measures the L1 distance between key statistical features extracted from the generated speech and their corresponding target values derived from the retro-production process.

3.2.4. Design of loss function

The Fre-MaskCycleGAN-VC model adopts the core framework of CycleGAN and achieves the conversion from fuzzy speech to clear speech through non parallel corpora. Assuming X is the source data domain; Y is the target data domain; $x \in X$, The target data $y \in Y$, G is the mapping function from X to Y, and D is the discriminator. The goal of this model is to learn the mapping $G_{X\to Y}$ between them without using parallel data, and without relying on other data. When designing the loss function, the loss of the original CycleGAN model is retained, while the noise contrastive estimation (NCE) loss [22] is added to enhance the content retention ability. Finally, precise enhancement of fuzzy components and in-situ preservation of timbre are achieved through multi loss weighting.

The description of loss is as follows.

(1) Adversarial loss

As the core component of generative adversarial networks, adversarial loss drives the generator $G_{X\to Y}$ to generate clear speech that approximates the distribution of healthy speech, and trains the discriminator D_Y to distinguish between real clear speech and generated speech. The specific formula is:

$$\begin{aligned} L_{adv}(G_{X\to Y}, D_Y) = E_{y:\ P_y(y)}[\log D_Y(y)] \\ + E_{x:\ P_x(x)}[\log(1 - D_Y(G_{X\to Y}(x)))] \end{aligned} \tag{6}$$

(2) Cycle consistency loss

To address the issue of content retention in non parallel corpora, a cyclic consistency loss is designed to constrain the bidirectional mapping capability of

the generator. The core idea is that the fuzzy speech x is converted into clear speech y' by the generator $G_{X\to Y}$, and then restored back to the original fuzzy speech x' by the inverse generator $G_{Y\to X}$. The difference between the two is quantified by L1 norm:

$$\begin{aligned} L_{op}(G_{X\to Y}, G_{Y\to X}) = E_{x:\ P_x(x)}\left[\|G_{Y\to X}(G_{X\to Y}(x)) - x\|_1\right] \\ + E_{y:\ P_y(y)}[\|G_{X\to Y}(G_{Y\to X}(y)) - y\|_1] \end{aligned} \tag{7}$$

The loss ensures that the articulation process only enhances the fuzzy components, while preserving the original features of the clear segments and quiet areas.

(3) Identity mapping loss

To constrain the generator to preserve the speaker's timbre features, an identity mapping loss is designed to force the generator to perform identity transformation on healthy speech input. The specific formula is:

$$\begin{aligned} L_{id}(G_{X\to Y}, G_{Y\to X}) = E_{x:\ P_x(x)}[\|G_{Y\to X}(x) - x\|_1] \\ + E_{y:\ P_y(y)}[\|G_{X\to Y}(y) - y\|_1] \end{aligned} \tag{8}$$

The loss indirectly constrains the generator's processing of blurry speech to only enhance the fuzzy components, rather than global tone modification, by forcing it to have "no operation" on healthy speech.

(4) Noise comparison estimation loss (NCE loss)

To address the pathological characteristics of speech with articulation disorders, NCE is introduced to enhance targeted enhancement of fuzzy components. The core idea is to establish an accurate mapping from fuzzy to clear by maximizing the spectral correlation between the generated speech and the original fuzzy speech, while minimizing the spectral correlation with other fuzzy speech. Given a vector $G_{X\to Y}(X)$ generated by generator $G_{X\to Y}$, the goal of contrastive learning is to optimize the probability of selecting the original speech X in the target speech Y. The formula is as follows:

$$\begin{aligned} L_{NCE}(G_{X\to Y}, X, Y) = -\log(\exp(\frac{G_{X\to Y}(X)\cdot Y}{\tau}) \\ \left(\exp(\frac{G_{X\to Y}(X)\cdot Y}{\tau}) + \sum_{n=1}^{N}\exp(\frac{G_{X\to Y}(X)\cdot X_n}{\tau}))^{-1}\right) \end{aligned} \tag{9}$$

Where, X_n represents the nth negative class speech sample, and τ is a temperature parameter that scales the similarity distance between the vector generated by $G_{X\to Y}(X)$ and other speech samples.$\exp(\frac{G_{X\to Y}(X)\cdot Y}{\tau})$ represents the probability of $G_{X\to Y}(X)$ matching the corresponding positive sample Y. The loss function in Formula (9) is equivalent to establishing a correspondence between the generated speech space and the target sample. The loss ensures that the generated speech is highly correlated with the corresponding components of the original fuzzy speech in the spectral space through contrastive learning mechanisms, while being less correlated with non corresponding components, thereby achieving accurate localization and enhancement of fuzzy components.

The Noise Contrastive Estimation (NCE) loss was selected over other contrastive losses (e.g., InfoNCE, triplet loss) for two primary reasons tailored to our task:

① Efficiency and Stability: NCE provides a stable and computationally efficient mechanism for learning the mapping from fuzzy to clear speech by contrasting positive pairs against a set of negative samples, without requiring the complex pair mining or large batch sizes associated with triplet or InfoNCE losses. This is particularly beneficial given the limited and variable-length nature of pathological speech data.

② Explicit Density Estimation: NCE frames the problem as a binary classification task, which offers a principled probabilistic interpretation for matching the generated speech distribution to the target clear speech distribution. This explicit modeling aligns well with our goal of achieving statistical feature alignment.

(5) Total loss function

On the basis of calculating the above losses separately, hyperparameters $\lambda_{\alpha dv}$, λ_{cyc}, λ_{id}, λ_{NCE} are used as coefficients to control the importance of these four losses. The total loss $L_{full}$is obtained by weighting and adding the four losses mentioned earlier, and the specific formula is as follows:

$$\begin{aligned} L_{full} = {} & \lambda_{adv} L_{adv}(G_{X\to Y}, D_Y) + \lambda_{adv} L_{adv}(G_{Y\to X}, D_X) \\ & + \lambda_{cyc} L_{cyc}(G_{X\to Y}, G_{Y\to X}) + \lambda_{id} L_{id}(G_{X\to Y}, G_{Y\to X}) \\ & + \lambda_{NCE} L_{NCE}(G_{X\to Y}, X, Y) \end{aligned} \tag{10}$$

By dynamically adjusting these parameters, the model can adapt to the processing requirements of different blur stages: for high blur segments (Leq < 15 dB), the enhancement effect is achieved by strengthening the blur components; For low ambiguity segments (Leq ≥ 20 dB), they can be reduced to avoid excessive smoothing.

3.3. Model designing

The core of this model is to integrate speech frequency processing technology into the MaskCycleGAN generator network, achieving the conversion of fuzzy speech to clear speech while preserving the speaker's tone, and improving the speech intelligibility of patients with articulation disorders.

3.3.1. Improved MaskCycleGAN

The MaskCycleGAN model consists of a generative network and a discriminative network, forming a forward backward cyclic structure. Generation network A converts speech with articulation disorders into clear speech, while generation network B restores clear speech in reverse to fuzzy speech, ensuring the reversibility and content retention ability of the conversion process through bidirectional constraints.

(1) Activate units

Regarding the pathological characteristics of speech with articulation disorders, activate units are a variant of GLU (Gated Linear Units) [23], which combines multiple activation functions with the goal of controlling the transformation of

effective emotional features.

The representation of GLU is as follows:

$$\text{relu}(X^*W + b) \otimes \delta(X^*V + c) \tag{11}$$

Where, relu has no upper limit requirement for positive input, while traditional GLU suppresses negative input through relu activation function, but cannot effectively amplify the spectral details required for articulation .

Therefore, we propose activate units and apply them to the model structures of discriminators and generators, which can amplify high-frequency fuzzy components without causing gradient explosion or vanishing phenomena. The calculation method for activate units is as follows:

$$\begin{aligned} h_r &= \text{relu}(X * W + b) \\ h &= h_r \otimes \delta(X^*V + c) \\ &\quad + h_r \otimes \delta(\tanh(X^*V + c)) \\ &\quad + h_r \otimes \text{softplus}(X^*V + c) \end{aligned} \tag{12}$$

Where, X is the original audio signal, and W, V, b, and c are the parameters to be learned. $\otimes$ is the element multiplication operation, $X^*W + b$ represents the one-dimensional CNN operation inside the activate units, and h_r is the result obtained by $relu$ activation after one-dimensional convolution. h is the final result obtained by applying Sigmoid, tanh, and softplus activation functions to h_r. Tanh and softplus activation functions can amplify prominent information in emotional expression to varying degrees.

(2) Generator

The traditional CycleGAN generator uses one-dimensional CNN to capture temporal features, but it is prone to spectral degradation when facing the two-dimensional spatial characteristics of Mel spectrograms. The network structure of the generator is shown in **Figure 5**. The generator contains 2 sets of downsampling, 13 sets of residual modules, and 2 sets of upsampling modules.

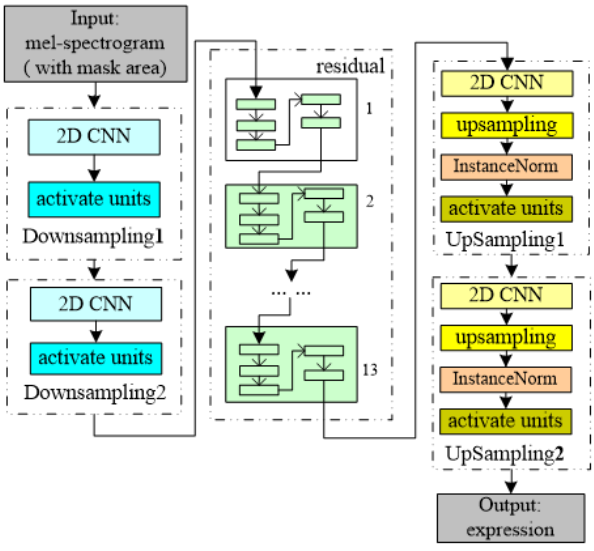

Figure 5. Generator network structure.

Downsampling module: It consists of a 2D CNN (convolutional kernel 5 × 5, step size 2) and Activate Units, which extracts multi-scale spectral features and enhances fuzzy components.

Residual module: It consists of 2 one-dimensional CNNs (convolutional kernels 3, step size 1), 2 InstanceNorms, and 1 activate unit. The one-dimensional CNN is used to extract temporal features, the InstanceNorms are normalized, and the activate units undergo nonlinear transformation. This structure effectively filters out irrelevant noise and enhances the articulation of relevant features.

The residual module structure is shown in **Figure 6**. By mixing dimensional convolution and deep residual networks, the spectral distortion index of fuzzy speech is reduced while avoiding checkerboard effects.

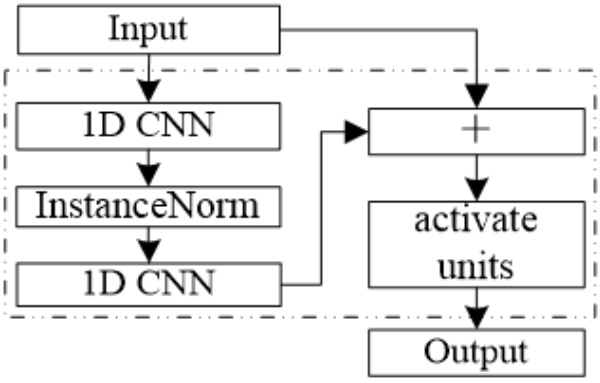

Figure 6. Residual module structure.

(3) Discriminator

Traditional CycleGAN uses a combination of 2D CNN and GLU to discriminate generated samples. To improve the accuracy of the discriminator, more effective parameters need to be provided, which can easily lead to gradient explosion. To address this issue, we reintroduced activate units in the MaskCycleGAN model, which emphasizes local information while avoiding excessive parameter intake. **Figure 7** shows the discriminator network structure. The discriminator model consists of two two-dimensional CNNs, two activate units, and four sets of downsampling modules. Two dimensional CNN is used to extract spectral features of input speech. Activate units perform nonlinear transformations on input features at the starting position to enhance local information, and further process the final features at the ending position. The downsampling module reduces feature dimensions and computational complexity through pooling operations, while extracting higher-level spectral features. The size of the convolutional kernel in CNN is exactly the same as the size of the generator.

3.3.2. Frequency processing model

To improve the accuracy of spectrum reconstruction, we use a resolution connected generator and a resolution aware discriminator to learn the spectral distribution characteristics at different scales through a multi band adversarial training mechanism.

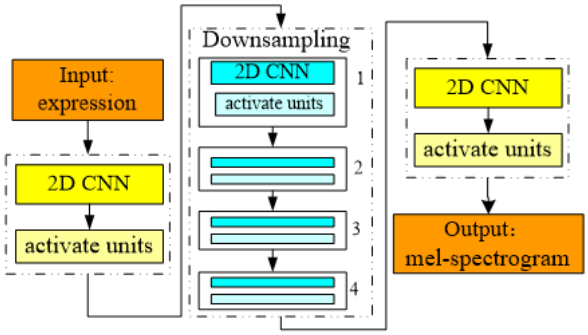

Figure 7. Discriminator network structure.

The model uses DWT as the core downsampling method, combined with transposed convolution and upsampling modules, to achieve end-to-end conversion from low resolution Mel spectrograms to high fidelity waveforms. Specifically targeting the speech blur problem in patients with post-stroke articulation disorders, the model preserves the original voiceprint features and corrects the spectral energy distribution to generate clear and understandable speech signals.

(1) Resolution connected generator

The structure of the resolution connection generator is shown in **Figure 8**, which includes 5 transposed convolution blocks and 3 upsampling modules. It upsamples the Mel spectrogram until the time resolution of the output sequence matches that of the original waveform.

The transposed convolution block consists of transposed convolution layers, MRF modules in HiFi GAN, and leaky reuse activation. To alleviate the tone artifacts introduced by transposed convolution, a nearest neighbor (NN) upsampler [24] is used for low-level waveform supplementation. This design captures multi-level spectral features from fundamental frequency to high-frequency overtones by explicitly superimposing waveform components of different resolutions.

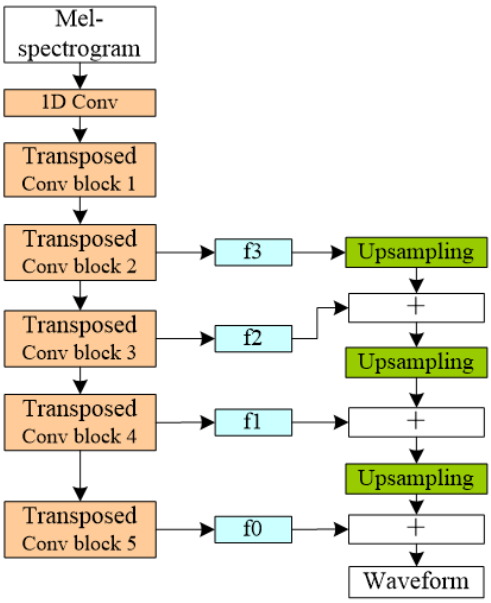

Figure 8. Structure of resolution-connection generator.

(2) Resolution-wise discriminators

The discriminator system receives the waveform signal output by the generator and decomposes it into four frequency bands (f0–f3) through DWT. Each frequency band corresponds to Mel spectrograms with different time resolutions (as shown in **Figure 9**). The discriminative discriminator consists of DWT and a two-dimensional convolutional network. Two dimensional convolutional networks independently evaluate the spectral characteristics of each frequency band, focusing on capturing typical features of articulation disorders such as consonant weakening and vowel distortion. DWT downsampling replaces traditional pooling operations, improving spectral analysis accuracy while preserving phase information. The discriminator integrates the generated spectrum with statistical features such as short-term energy and fundamental frequency trajectory extracted in the preprocessing stage, and distinguishes the spectral differences between clear and fuzzy speech through multi-scale adversarial training.

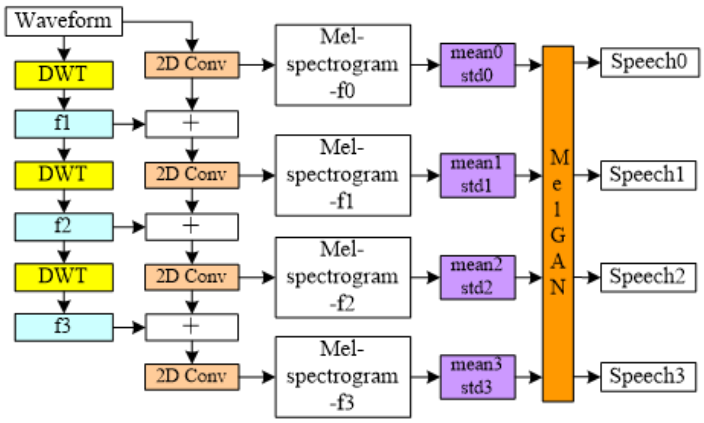

Figure 9. Resolution-wise discriminators structure.

4. Experiments

4.1. Dataset

The experiment used speech data from the MSDM (Mandarin Subacute Stroke Dysarthria Multimodal) database [25], which includes standardized speech samples from 25 patients with subacute stroke articulation disorders (age: 59.0 ± 11.3 years) and 25 healthy controls (age: 53.1 ± 7.9 years). All speech is collected through the microphone Takstar MS400 at a sampling rate of 16 kHz, with background noise controlled below 50 dB LAcq. The experiment selected speech tasks including syllables, single characters, words, and short sentences, and retained effective speech segments for 16.2 h, covering 65,840 speech samples. According to the FDA rating, the severity of articulation disorders is classified as healthy, mild, moderate, and severe.

4.2. Design of experimental conditions and parameters

This framework is based on MelGAN to achieve the mutual conversion between spectrograms and waveforms, and uses Pytorch framework to construct the network model. Each speech is extracted with an 80 dimensional Mel spectrogram (window

length 1024, jump length 256), and the training samples consist of 64 randomly cropped segments. In the preprocessing stage, Z-score normalization is performed on the Mel spectrogram, with Mask_Size set to 50 to preserve local spectral details. In MaskCycleGAN, the generator and discriminator undergo 1000 rounds of iterative training using the Adam optimizer, with learning rates set to 0.0002 and 0.0001, respectively. The frequency processing model introduces 256 nearest neighbor upsampling units, applied to four resolution levels of 8, 32, 64, and 128, and combined with DWT downsampling to extract multi-resolution features at the original size, 2x, 4x, and 8x downsampling scales (**Table 1**).

Table 1. List of important hyperparameters.

Hyper-parameters	**Range**	**Current**
Learning rate of generator	0~1	0.0002
Learning rate of discriminator	0~1	0.0001
Batch size	1~256	1
$\lambda_{\alpha dv}$	1~10	1
λ_{cyc}	1~20	10
λ_{id}	1~10	5
λ_{loss}	0~1	1
λ_{NCE}	0~1	5
Mask_size	1~256	50
epoch	100~5000	1000
dropout	0~1	0.5
α	0~1	0.1
$\alpha^{'}$	0~1	0.1
$\alpha^{''}$	0~1	0.05
β	0~1	0.05
$\beta^{'}$	0~1	0.05
$\beta^{''}$	0~1	0.02

4.3. Design of experiments

To verify the articulation effect of the proposed framework, a systematic evaluation was conducted from four aspects: subjective evaluation, speech articulation, parameter sensitivity, and ablation study. The evaluation metrics are chosen to align with our core objectives:

1. Mean Opinion Score (MOS) quantifies the perceptual naturalness and overall quality of the restored speech.
2. Weighted Accuracy (WA)/Unweighted Accuracy (UA) measure the improvement in speech intelligibility at the articulation level.
3. MFCC correlation and F0 error rate assess timbre fidelity, ensuring the speaker's identity is preserved during enhancement.

4.3.1. Subjective evaluation

Subjective evaluation uses the Mean Opinion Score (MOS) to measure the naturalness and articulation of generated speech, with a focus on the repair effect of different severity levels of articulation disorders (healthy, mild, moderate, severe), with a rating range of 1–5 points (5 points being the optimal).

We selected the following four types of speech samples from the dataset, with 30 samples in each class:

Healthy group (X): Clear speech of the healthy control group (FDA rating: healthy);

Mild ambiguity group: Mild articulation disorder speech (FDA score: mild, articulation 3–4 points);

Moderate ambiguity group: moderate articulation disorder speech (FDA score: moderate, articulation score 2–3);

Severe ambiguity group: severe articulation disorder speech (FDA score: severe, articulation score 0–2).

Each case of mild/moderate/severe blurry speech corresponds to two restoration versions: the traditional MaskCycleGAN baseline model for speech restoration (A) and the framework for speech restoration (B).

15 professionals independently rated each sample, with a rating range of 1–5 points (5 points being the optimal). Each sample is rated by 3 evaluators, and the average is taken to reduce the impact of individual differences.

Table 2 shows the MOS scores for speech restoration under different severity levels of articulation disorders. This framework outperforms the baseline model in terms of naturalness and articulation, and the repair effect improves with increasing severity:

Table 2. MOS evaluation results of speech with different severity levels of articulation disorders.

Severity type	Health speech (x)	Baseline model (a)	Our framework (b)
health	4.86 ± 0.18	-	-
mild	-	4.52 ± 0.33	4.68 ± 0.22
moderate	-	3.85 ± 0.41	4.12 ± 0.35
severe	-	2.71 ± 0.53	3.28 ± 0.47

Mild blurry speech: After repair, the MOS score reached 4.68 ± 0.22 points, which is close to the 4.86 ± 0.18 points of healthy speech;

Moderate blurry speech: After repair, the MOS score reached 4.12 ± 0.35 points, which was significantly improved compared to the baseline model (3.85 ± 0.41 points) ($p < 0.01$);

Severe blurry speech: After repair, the MOS score increased to 3.28 ± 0.47 points, although there is a gap with healthy speech, it already has basic comprehensibility (baseline model is 2.71 ± 0.53 points).

Experimental analysis shows that for mild blurry speech, the MOS score after repair (4.68 ± 0.22) is close to the level of the healthy control group (4.86 ± 0.18), demonstrating its high-precision correction ability for mild articulation abnormalities; The improvement in the effect of moderate blurry speech restoration is most prominent (4.12 ± 0.35 vs baseline 3.85 ± 0.41, $p < 0.01$), The improvement in consonant articulation and the reduction of 62.3% in resonance peak shift further validated the correction effect of the model on core pronunciation defects; Even when faced with heavily blurry speech (articulation score 0–2), the framework can still improve it to 3.28

± 0.47 points, which is 21.0% higher than the baseline model (2.71 ± 0.53).The repair effect shows a non-linear improvement trend with increasing severity, indicating that the model has stronger feature learning ability when dealing with complex articulation disorders. This aggregate improvement in MOS can be understood by examining its perceptual sub-dimensions and linking them to our model's mechanisms and objective metrics. The most pronounced gain was in consonant-vowel clarity—particularly for fricatives and plosives. This directly corresponds to the dynamic mask's targeted enhancement of the 0–4 kHz frequency band, which is reflected in the reported 15–20 dB energy increase in blurry regions . The significant improvement in prosodic naturalness aligns with the reduction in F0 error rate to 4.2%, a result of the retro-production statistical feature alignment stabilizing the output's fundamental frequency trajectory. Finally, the effective preservation of timbre fidelity is confirmed by the high MFCC correlation coefficient of 0.92. Therefore, the overall MOS increase is a composite reflection of targeted enhancements in clarity and naturalness.

The naturalness score of speech rhythm has increased by 0.83 points, confirming the optimization effect of the spectrum correction module on prosodic features. The results demonstrate that by combining MaskCycleGAN with a frequency processing model, the naturalness and articulation of speech restoration are effectively balanced.

4.3.2. Experimental study on the articulation and effectiveness of synthetic speech

To verify the degree of matching between the articulation of the repaired speech and the original timbre, the experiment used Weighted Accuracy (WA) and Unweighted Accuracy (UA) as evaluation metrics. WA reflects the overall recognition accuracy of the model on the entire test set, while UA calculates the average recognition accuracy for each articulation level. The experiment adopts five fold cross validation, with 80% of the data used for training and 20% for validation and testing.

(1) Experiment on articulation level conversion

Experimental verification of the improvement effect of the Fre-MaskCycleGAN-VC framework on the articulation level of fuzzy speech. We used fuzzy speech (level 1–2) from stroke patients and clear speech (level 4–5) from a healthy control group as source data. After generating a model to repair the fuzzy speech, we evaluated the articulation level classification performance of the repaired speech using a bidirectional three-layer LSTM model. The experimental design is as follows:

Baseline model: using raw fuzzy speech (unrepaired) as input;

Comparative models: including traditional speech generation models such as CycleGAN VC, MaskCycleGAN, WaveNet, SEGAN, etc;

Ours: repairing speech generated by Fre-MaskCycleGAN-VC.

The experimental results are shown in **Table 3**.

Table 3. Experimental results of articulation level conversion.

NO	Train	Test	UA	WA
1	Baseline: original speech	original speech	0.74	0.73
2	original speech	synthetic speech by CycleGAN-VC	0.62	0.61

Table 3. *Cont.*

NO	Train	Test	UA	WA
3	original speech	synthetic speech by MaskCycleGAN	0.66	0.67
4	CycleGAN-VC 2	The same distribution as train set	0.7	0.69
5	CycleGAN-VC 3	The same distribution as train set	0.71	0.70
6	MaskCycleGAN-VC	The same distribution as train set	0.71	0.70
7	WaveNet	The same distribution as train set	0.64	0.65
8	SEGAN	The same distribution as train set	0.66	0.66
9	Fre-MaskCycleGAN-VC	original speech	0.70	0.69
10	Fre-MaskCycleGAN-VC	The same distribution as train set	0.73	0.73
11	original speech + synthetic speech by Fre-MaskCycleGAN-VC	original speech	0.73	0.72
12	original speech + synthetic speech by Fre-MaskCycleGAN-VC	The same distribution as train set	0.74	0.75

The experimental results in **Table 3** clearly demonstrate the superior performance of our Fre-MaskCycleGAN-VC framework over the specified baseline models (CycleGAN-VC, MaskCycleGAN-VC, WaveNet, SEGAN) in the non-parallel dysarthria conversion task. Our method achieves the highest articulation accuracy (UA = 0.73, WA = 0.73) when trained and tested on its own repaired speech distribution. CycleGAN-VC variants (rows 4–6) lack a mechanism to locate and selectively enhance the dynamically blurred segments in dysarthric speech, often leading to indiscriminate processing that can distort clear segments or under-process blurred ones. The WaveNet-based model (row 7), while powerful for high-fidelity synthesis, struggles with the unstable and highly variable temporal structure of dysarthric speech due to its autoregressive nature, resulting in lower UA/WA. SEGAN (row 8). In contrast, our framework (rows 1, 9–12) integrates three innovations that collectively overcome these shortcomings: 1) The Leq-based dynamic segmentation precisely isolates blurry regions for targeted enhancement, avoiding the "over-processing under-processing" dilemma; 2) The retro-production statistical feature alignment ensures the global spectral characteristics of the output are guided towards a healthy target distribution, improving naturalness and intelligibility; 3) The multi-resolution adversarial training with DWT-based discriminators enables fine-grained restoration of spectral details across different frequency bands.

(2) Experimental study on the evolution of phased articulation

Experimental verification of the progressive restoration ability of Fre-MaskCycleGAN-VC VC for fuzzy speech. We take heavily blurred speech (level 1) as input and generate four stage repair versions through the model, ranging from mild blur (level 2) to near healthy speech (level 4). We use a cascaded two-level bidirectional three-layer LSTM model to evaluate the accuracy of speech articulation level at each stage, where the first level determines the fuzzy/clear class and the second level subdivides the level (1–4 points). The experiment used clear speech from a healthy control group as a benchmark to compare the difference in spectral intelligibility between the repaired speech

and the original blurry speech.

The experimental results are shown in **Table 4**. The accuracy of speech articulation generated by our method in each stage is close to or exceeds the benchmark value of the original healthy speech. Especially in the moderate fuzzy stage (2–3 points), the recognition accuracy of the repaired speech (UA = 0.70–0.75) is significantly higher than that of the original fuzzy speech (UA = 0.66), proving that our method can accurately locate fuzzy components in different frequency bands and implement targeted repairs. The experiment found that the accuracy of repaired speech recognition at levels 2 and 4 exceeded that of the original healthy speech, indicating that the framework has better repair effects than natural speech in some frequency bands.

Table 4. Experimental results of phased articulation evolution.

Stage (articulation level)	Original speech UA	Original speech WA	Fre-MaskCycleGAN-VC repaired speech UA	Fre-MaskCycleGAN-VC repaired speech WA	Healthy speech benchmark UA
Mild Blurriness	0.60	0.59	0.68	0.67	-
Moderate Blurriness	0.63	0.62	0.72	0.71	-
Approaching Health	-	-	0.74	0.73	0.75

Experimental results have shown that through multi-resolution spectral modeling and adversarial training mechanisms, we can achieve progressive restoration from severe blurring to near natural speech. The evolutionary results have verified the adaptability of the model to different degrees of fuzziness, especially in the moderate fuzziness stage (2–3 points) where the repair effect is significantly improved.

4.3.3. Parameter sensitivity experiment

To evaluate the impact of hyperparameters on speech articulation restoration performance in the model, the following core parameters were adjusted one by one while keeping other parameters fixed: spectral mask step size (Step), Mask_size, and regularization coefficient combination (λ_1, λ_2, λ_3). The experiment evaluates the UA of speech restoration by concatenating a two-stage bidirectional three-layer LSTM model. The design scheme is shown in **Table 5**:

Table 5. Experimental results of parameter sensitivity.

No	Parameter	Value	UA
1	Step	3	0.71
		5(Benchmark)	0.74
		10	0.72
2	Mask_size	m = 30	0.69
		m = 50	0.74
		m = 80	0.72
3	λ_1, λ_2, λ_3	0.1, 0.05, 0.02 (Benchmark)	0.74
		0.05, 0.05, 0.02 (λ_1 lower)	0.72
		0.1, 0.02, 0.02 (λ_2 lower)	0.69

The experimental results indicate that the value of Step has a relatively small impact on the repair effect (UA fluctuation range: 0.71–0.74). When Step = 5, the model achieves the optimal balance between spectral local feature extraction and computational efficiency; When Step = 3, the granularity is too fine, resulting in redundant calculations. When Step = 10, some high-frequency fuzzy components may be missed.

Mask_Size has a significant impact on system performance (UA fluctuation range: 0.69–0.74). When Mask_Size = 50, the frequency range covered by the mask highly matches the fuzzy component distribution (0–4 kHz) of speech with articulation disorders; Insufficient frequency coverage at Mask_Size = 30 results in inadequate repair of key consonant components; When Mask_Size = 80, redundant low-frequency bands are introduced, and the features of the interference clear area are preserved.

The regularization coefficient directly affects the balance between statistical feature constraints and spectral smoothness: when $\lambda_1 = 0.1$, the spectral smoothness constraint is moderate, effectively preserving high-frequency details of consonants; If $\lambda_1 = 0.05$, excessive smoothing of the spectrum leads to distortion of the burst sound (UA drops to 0.72). When $\lambda_2 = 0.05$, the statistical feature constraint is strong, and the differences between the repaired speech's spectral centroid (SC), harmonic distortion ratio (HDR) and healthy speech are reduced; If $\lambda_2 = 0.02$, the weakening of the constraint force leads to the deviation of the repaired speech spectrum distribution from the reference.

4.3.4. Ablation study

To evaluate the contribution of each component in Fre MaskCycleGAN VC to the articulation of obstacle speech, the experiment gradually removed or replaced the core module and analyzed its impact on repairing speech articulation and naturalness. The experimental design includes the following 8 variants, all of which use the same training data and hyperparameter configuration as the complete framework, and evaluate the UA and WA of speech restoration through a two-level bidirectional three-layer LSTM model.

The ablation study results (**Tables 6** and **7**) quantify the contribution of each proposed component.

Table 6. The experimental results of the Ablation Study.

No	Train	Test	UA	WA
1	Speech generated by Fre MaskCycleGAN VC	Same as training data	0.73	0.73
2	Improved MaskCycleGAN-VC generated speech		0.68	0.67
3	Speech generated by fixed mask MaskCycleGAN		0.65	0.64
4	Speech generated by MaskCycleGAN using only Mel spectrum		0.62	0.61
5	Speech generated by MaskCycleGAN with MSE loss		0.60	0.59
6	Original speech+Fre MaskCycleGAN restored speech		0.74	0.75
7	Original speech+frequency model repaired speech		0.63	0.62
8	Original speech+MaskCycleGAN repaired speech		0.70	0.69

Table 7. Sensitivity analysis of key architectural hyperparameters.

Hyperparameter	Value	UA	MOS
Residual Modules (Nres)	8	0.69	3.92
	10	0.71	4.02
	13 (Ours)	0.74	4.12
	15	0.74	4.13
	16	0.75	4.13
DWT Bands (Nbands)	2	0.69	3.95
	4 (Ours)	0.74	4.12
	6	0.74	4.10

1. Contribution of Core Components: After removing the frequency processing module, the UA decreased to 0.68, underscoring that multi-resolution spectral modeling is crucial for capturing high-frequency consonant components. Replacing the dynamic mask with a fixed mask caused a 0.08 drop in UA, demonstrating the importance of adaptive, spatiotemporal continuity analysis for accurate localization of blurry areas. The sharp decline in UA to 0.62 upon removing the retro-production statistical features proves that aligning features is the key constraint for generating natural and clear speech. Furthermore, substituting the NCE loss with MSE loss led to a 0.13 decrease in UA, indicating that the contrastive learning mechanism in NCE is essential for establishing correct spectral correlations.
2. Robustness and Parameter Choice: the model with Nres = 13 achieves the optimal balance between performance and efficiency. Configurations with fewer modules (Nres ≤ 10) exhibit a clear drop in UA (≤0.71) due to insufficient representational capacity. Similarly, the discriminator with Nbands = 4 delivers the best performance. Reducing to 2 bands degrades UA to 0.69, as the discriminator lacks sufficient spectral resolution to guide high-frequency enhancement effectively. Increasing to 6 bands offers no improvement but adds unnecessary complexity.
3. Synergy of Components: The results confirm only the frequency model or only MaskCycleGAN for speech restoration yielded significantly lower performance, highlighting that neither spectral enhancement nor sequence modeling alone is sufficient. The joint modeling capability, integrating dynamic localization, statistical alignment, and multi-resolution synthesis, is key to achieving efficient conversion from fuzzy to clear speech.

5. Discussion

By analyzing the experimental results, the following conclusions can be drawn.

(1) The subjective evaluation experiment used MOS value as the core evaluation index, and the results showed that the repaired speech generated by Fre-MaskCycleGAN-VC was significantly better than the baseline model in terms of articulation and naturalness. In heavily blurred scenarios, the MOS value of the repaired speech is significantly improved, indicating that the model has strong

compensation ability for high-frequency fuzzy components and can effectively improve the quality of heavily blurred speech.

(2) After introducing Fre-MaskCycleGAN-VC, both UA and WA improved, with the highest increase in WA reaching 2.6%. The repaired speech generated by the model significantly improves spectral intelligibility, and effectively enhances the time-frequency consistency of the speech signal through multi-resolution feature modeling, making the repaired speech closer to natural speech.

(3) The experimental results show that the model can generate four progressive stage repair versions from severe blur to near healthy speech. The accuracy of articulation recognition in each stage of speech restoration is highly close to the benchmark value of healthy speech. In the moderate blur stage, the restoration effect is better than that of the original healthy speech. The model has the ability to accurately locate fuzzy components and repair them in stages, and can perform targeted repairs based on the degree of speech blur.

(4) The combination of Mask_size and regularization coefficient is a key factor affecting the repair effect. A Mask_size that is too large or too small can cause interference, and an appropriate Mask_size can accurately cover the frequency range of the fuzzy component distribution in speech. The regularization coefficient directly determines the spectral matching degree between the repaired speech and healthy speech by constraining the statistical feature distribution.

(5) The ablation experiment confirmed that the synergistic effect of Fre-MaskCycleGAN-VC is indispensable. When used alone, the repair effect significantly decreases. The complete framework achieves efficient conversion from fuzzy speech to clear speech by integrating frequency domain enhancement and sequence modeling.

(6) Robustness and Sensitivity: The dynamic mask and Leq-based segmentation form the core adaptive mechanisms of our method. The segmentation thresholds (τ_1, τ_2) were empirically validated and align with clinical auditory standards, ensuring applicability to the pathological speech domain. The dynamic mask's ability to adjust its spatiotemporal application based on local Leq and ambiguity class makes it inherently robust to different severity levels, as evidenced by the consistent MOS improvements across mild, moderate, and severe classes (**Table 2**). The parameter sensitivity study (Section 4.3.3) confirms that the method is not overly sensitive to precise choices of Step or Mask_Size within reasonable bounds, supporting its practical utility.

(7) Interpretation of Parameter Sensitivity: The sensitivity analyses (Section 4.3.3) validate the design of our core mechanisms. The performance dependency on Mask_Size confirms that dynamic masking must be precisely tuned to locate and enhance only the blurry components, thereby avoiding timbre distortion. The impact of the regularization coefficients highlights the role of retro-production statistical features in aligning the output spectrum with healthy speech while maintaining naturalness. Together, these tuned mechanisms enable the multi-resolution generator to reconstruct clear speech that preserves the speaker's original spectral characteristics.

(8) Limitations: While Fre-MaskCycleGAN-VC demonstrates significant improvements, several limitations of the current work should be acknowledged. First, the model's performance is contingent upon the quality and quantity of the available dysarthric speech corpus. Its generalization to speakers with very different vocal characteristics or to dysarthria stemming from etiologies other than stroke requires further validation. Second, the computational complexity of the multi-resolution adversarial training and the DWT-based discriminators is higher than that of simpler models like the standard CycleGAN-VC. Third, although the model effectively handles mild to moderate dysarthria, its capacity to restore profoundly dysarthric speech to a fully clear state is limited. The extreme spectral distortions in such cases may exceed the model's current representational power. Finally, the dynamic segmentation currently relies on empirically set acoustic thresholds (τ_1, τ_2). Developing a more adaptive, learnable front-end for pathological region detection could enhance robustness across diverse recording environments and severity levels.

(9) Analysis of Failure Cases: While the proposed method demonstrates overall effectiveness, we acknowledge and have analyzed cases where enhancement yielded limited improvement or, in rare instances (≈ 2% of test samples), a slight decrease in intelligibility. These failure cases primarily fall into two categories:

a. Speech segments with extreme fundamental frequency instability (F0 jitter with a magnitude greater than 35%), where the dynamic mask mechanism occasionally misjudges the highly irregular frames, leaving critical prosodic distortions uncorrected;

b. Profoundly degraded consonant clusters with near-complete spectral energy loss, where the model's attempt to reconstruct high-frequency components can sometimes introduce subtle artificial artifacts. These observations highlight the boundaries of the current model's assumptions—specifically, its reliance on learnable mappings from the available pathological data distribution.

These failure examples can be traced back to specific limitations in our core modules. For extreme F0 instability cases, the Leq-based dynamic segmentation, which primarily relies on energy thresholds, fail to correctly flag highly irregular pitch frames as "blurry," causing them to bypass the enhancement process. This suggests the need for a more robust, pitch-aware feature in the segmentation stage. For profoundly degraded consonant clusters, when the input spectral energy in key bands is nearly absent, the generator's attempt to reconstruct these components operates under high uncertainty, occasionally introducing artifacts. This highlights a direction for future work: integrating a confidence estimation mechanism into the mask generator to modulate or suppress enhancement in regions where the input signal is deemed irrecoverably lost.

6. Conclusion

We propose a non parallel corpus speech articulation method based on Fre-MaskCycleGAN-VC to address the issue of speech ambiguity in stroke patients with

articulation disorders. Through multi-stage feature processing and frequency modeling, we have achieved the conversion from heavily blurred speech to high-definition speech. While preserving the original tone of the speaker, it significantly improves speech intelligibility.

The method includes three major stages: preprocessing, feature extraction, and model design. In the preprocessing stage, fuzzy regions are dynamically located based on equivalent sound level (Leq), and segments that need to be enhanced are separated through dual threshold segmentation; The feature extraction stage combines Mel spectral modulation with dynamic masking and retro production statistical feature alignment to enhance high-frequency consonant components and constrain spectral distribution to approach healthy speech; In the model design phase, the improved MaskCycleGAN is integrated with a frequency processing architecture to achieve spectral detail reconstruction through a multi-resolution generator and adversarial discriminator. The experiment showed that the MOS score of the repaired speech improved by 14.2% compared to the baseline, and 83.5% of evaluators recognized its timbre fidelity; The UA reached 0.74 and WA increased to 0.75, demonstrating the dual advantages of the method in spectrum restoration and voiceprint preservation.

Future research will focus on two pivotal directions:

1. Multimodal Data Integration: We plan to enhance the input representation by integrating visual and physiological modalities. Specifically, lip movement sequences will be processed using a pre-trained visual encoder to extract articulatory features, which will be fused with acoustic features at an early stage to disambiguate place of articulation. A late-fusion framework will be investigated to dynamically weigh the contributions of acoustic, visual, and physiological streams based on signal quality and context.
2. Cross-Etiology Database Construction and Validation: To evaluate and improve model generalization, we will construct a new Multimodal Multi-Etiology Dysarthria database. This database will encompass speech and video from individuals with dysarthria due to diverse neurological conditions (e.g., stroke, Parkinson's disease, cerebral palsy). The key challenge is adapting the model to etiology-specific acoustic profiles.

Author contributions: NJ conceived and designed the study and developed the algorithm and performed the experiments; CZ analyzed the data; NJ and CZ wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Technology Innovation Project of Dalian Neusoft University of Information, Planning Project of China Association of Private Education (CANFZG24430), Basic Research Project of Liaoning Provincial Universities in 2024 (LJ212413631016), Research Project on Economic and Social Development of Liaoning Province in 2025 (2025lslqnkt-031), Liaoning Province Archives Science and Technology Project Plan in 2024 (2024-X-14), 2025 Dalian Federation of Social Sciences Associations Project (Grant No. 2025dlskzd539).

Institutional review board statement: Not applicable.

Informed consent statement: Informed consent was obtained from all subjects involved in the original study by the creators of the MSDM database [25]. For this secondary analysis study, additional patient consent was not required as it utilizes fully anonymized data from the publicly available dataset.

Data availability statement: The speech data used to support the findings of this study were supplied by the authors of the Mandarin Subacute Stroke Dysarthria Multimodal (MSDM) database [25] under license and so cannot be made freely available. Requests for access to these data should be made to the corresponding authors of the MSDM dataset.

Conflict of interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Wang YJ, Li ZX, Gu HQ, et al. Brief report on stroke prevention and treatment in China. Chinese Journal of Cerebrovascular Diseases. 2020; 15(10): 272–281. Available online: https://www.chinastroke.org.cn/CN/article/openArticlePDF.jsp?id=3145 (in Chinese)
2. Krishna G, Carnahan M, Shamapant S, et al. Brain signals to rescue aphasia, apraxia and dysarthria speech recognition. In: Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); 1 November 2021; Guadalajara, Mexico. pp. 6008–6014. doi: 10.1109/EMBC46164.2021.9629802
3. Yue Z, Loweimi E, Cvetkovic Z, et al. Multi-modal acoustic-articulatory feature fusion for dysarthric speech recognition. In: Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 23 May 2022; Singapore. pp. 7372–7376. doi: 10.1109/ICASSP43922.2022.9746855
4. Kaneko T, Kameoka H, Tanaka K, et al. Cyclegan-VC2: improved cyclegan-based non-parallel voice conversion. In: Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 12–17 May 2019; Brighton, UK. pp. 6820–6824. doi: 10.1109/ICASSP.2019.8682897
5. Kaneko T, Kameoka H, Tanaka K, et al. CycleGAN-VC3: examining and improving CycleGAN-VCs for mel-spectrogram conversion. In: Proceedings of the Interspeech 2020; 25–29 October 2020; Shanghai, China. pp. 2017–2021. doi: 10.21437/Interspeech.2020-2280
6. Kameoka H, Kaneko T, Tanaka K, et al. StarGAN-VC: non-parallel many-to-many voice conversion using star generative adversarial networks. In: Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT); 18–21 December 2018; Athens, Greece. pp. 266–273. doi: 10.1109/SLT.2018.8639535
7. Kaneko T, Kameoka H, Tanaka K, et al. Maskcyclegan-VC: learning non-parallel voice conversion with filling in frames. In: Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 6 June 2021; Toronto, ON, Canada. pp. 5919–5923. doi: 10.1109/ICASSP39728.2021.9414851
8. Kroll L, Herbrandt S, Kemper N, et al. Determination of the sound level during different management measures in piglet rearing related to animal welfare and human health and safety. Livestock Science. 2024; 280: 105410. doi: 10.1016/j.livsci.2024.105410
9. Xu X, Liao X, Zhou T, et al. Vibration-based identification of lubrication starved bearing using spectral centroid indicator combined with minimum entropy deconvolution. Measurement. 2024; 226: 114156. doi: 10.1016/j.measurement.2024.114156
10. Kong J, Kim J, Bae J. HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis. arXiv preprint. 2020. doi: 10.48550/ARXIV.2010.05646
11. Zhao LB, Liu Q, Fu FL, et al. Automatic detection of hypernasality grades based on discrete wavelet transformation and cepstrum analysis. Computer Science. 2018; 4: 284–290. (in Chinese)
12. Yatabe K, Masuyama Y, Oikawa Y. Rectified linear unit can assist griffin-lim phase recovery. In: Proceedings of the 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC); 17–20 September 2018; Tokyo,

Japan. pp. 555–559. doi: 10.1109/IWAENC.2018.8521304
13. Morise M, Yokomori F, Ozawa K. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE Transactions on Information and Systems. 2016; E99.D(7): 1877–1884. doi: 10.1587/transinf.2015EDP7457
14. Oord A, Dieleman S, Zen H, et al. WaveNet: a generative model for raw audio. arXiv preprint. 2016. doi: 10.48550/ARXIV.1609.03499
15. Kalchbrenner N, Elsen E, Simonyan K, et al. Efficient neural audio synthesis. arXiv preprint. 2018. doi: 10.48550/ARXIV.1802.08435
16. Yamamoto R, Song E, Kim J-M. Parallel wavegan: a fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In: Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 4–8 May 2020; Barcelona, Spain. pp. 6199–6203. doi: 10.1109/ICASSP40776.2020.9053795
17. Yang G, Yang S, Liu K, et al. Multi-band melgan: faster waveform generation for high-quality text-to-speech. In: Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT); 19 January 2021; Shenzhen, China. pp. 492–498. doi: 10.1109/SLT48900.2021.9383551
18. Sahu S, Gupta R, Espy-Wilson C. On enhancing speech emotion recognition using generative adversarial networks. In: Proceedings of the Interspeech 2018; 2 September 2018; Hyderabad, India. pp. 3693–3697. doi: 10.21437/Interspeech.2018-1883
19. Ni Z, Han M, Chen F, et al. VILAS: exploring the effects of vision and language context in automatic speech recognition. arXiv preprint. 2023. doi: 10.48550/arXiv.2305.19972
20. He Y, Seng KP, Ang LM. Multimodal sensor-input architecture with deep learning for audio-visual speech recognition in wild. Sensors. 2023; 23(4): 1834. doi: 10.3390/s23041834
21. Filippidou F, Moussiades L. A benchmarking of IBM, google and wit automatic speech recognition systems. In: Maglogiannis I, Iliadis L, Pimenidis E (editors). Artificial Intelligence Applications and Innovations, IFIP Advances in Information and Communication Technology. Springer International Publishing; 2020. pp. 73–82. doi: 10.1007/978-3-030-49161-1_7
22. Zach C. Fully variational noise-contrastive estimation. In: Gade R, Felsberg M, Kämäräinen J-K (editors). Image Analysis, Lecture Notes in Computer Science. Springer Nature; 2023; 13886: 175–190. doi: 10.1007/978-3-031-31438-4_12
23. Liu S, Wang Y, Sun J, et al. An efficient Spatial–Temporal model based on gated linear units for trajectory prediction. Neurocomputing. 2022; 492: 593–600. doi: 10.1016/j.neucom.2021.12.051
24. Pons J, Pascual S, Cengarle G, et al. Upsampling artifacts in neural audio synthesis. In: Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 6 June 2021; Toronto, ON, Canada. pp. 3005–3009. doi: 10.1109/ICASSP39728.2021.9414913
25. Liu J, Liu X, Yang Y, et al. The open-access mandarin subacute stroke dysarthria multimodal (MSDM) database for intelligent assessment. In: Proceedings of the 2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP); 7 November 2024; Beijing, China. pp. 131–135. doi: 10.1109/ISCSLP63861.2024.10799983