Article

Not peer-reviewed version

MPFM-VC: A Voice Conversion Algorithm based on Multi-Dimensional Perception Flow Matching

Yanze Wang *, Xuming Han, Shuai Lv, Ting Zhou, Yali Chu

Posted Date: 18 April 2025

doi: 10.20944/preprints202504.1428.v2

Keywords: voice conversion; flow matching; multi-dimensional; content feature perturbation; adversarial training

Article

MPFM-VC: A Voice Conversion Algorithm based on Multi-Dimensional Perception Flow Matching

Yanze Wang [1,*], Xuming Han [1], Shuai Lv [1], Ting Zhou [1] and Yali Chu [2]

[1] School of Information Science and Technology, Jinan University, Guangzhou, Guangdong, 510632, China
[2] School of Mathematics and Statistics, Changchun University of Technology, Changchun, Jilin, 130000, China
* Correspondence: hanxibird@gmail.com

Abstract: Voice conversion (VC) is a cutting-edge technology that enables the transformation of raw speech into high-quality audio resembling the target speaker's voice while preserving the original linguistic content and prosodic patterns. In this study, we propose a novel voice conversion algorithm, Multi-Dimensional Perception Flow Matching (MPFM-VC). Unlike traditional approaches that directly generate waveform outputs, MPFM-VC models the evolutionary trajectory of mel spectrograms with a flow matching framework and incorporates a multi-dimensional feature perception network to enhance the stability and quality of speech synthesis. Additionally, we introduce a content perturbation method during training to improve the model's generalization ability and reduce inference-time artifacts. To further increase speaker similarity, an adversarial training mechanism on speaker embeddings is employed to achieve effective disentanglement between content and speaker identity representations, thereby enhancing the timbre consistency of the converted speech. Experimental results for both speech and singing voice conversion tasks demonstrate that MPFM-VC outperforms existing state-of-the-art VC models in both subjective and objective evaluation metrics. The synthesized speech exhibits significantly improved naturalness, clarity, and timbre fidelity, validating the effectiveness of the proposed approach.

Keywords: voice conversion; flow matching; multi-dimensional; content feature perturbation; adversarial training

1. Introduction

Voice conversion (VC) [1] is an advanced speech processing technique that transforms a source speaker's voice into that of a target speaker while preserving the original linguistic content and prosody. This technology has gained significant traction in applications such as personalized speech synthesis, voice editing for film and television, and speech anonymization.

Recent advances in voice conversion (VC) have largely been built upon the development of text-to-speech (TTS) frameworks [2–6]. Current VC models can be broadly classified into two categories. The first category includes end-to-end systems[7–9], which directly map input speech to output speech within a unified architecture. These systems typically outperform traditional models[10] in terms of speaker similarity and robustness. The second category follows a cascaded architecture, in which automatic speech recognition (ASR) models [11–13] are used to extract content features; then, these characteristics are transferred to the acoustic model based on the generated neural network [14–18] to produce a mel spectrogram, which is subsequently converted into a waveform [19–21] by a neural vocoder. While cascaded models often achieve superior audio quality compared with end-to-end systems, they are generally more vulnerable to noise and less robust under domain-shift conditions.

Despite progress in the field, several challenges remain, particularly in real-world applications. First, VC systems are highly sensitive to noise, especially when handling low-quality or degraded speech inputs. Second, distributional shifts between the training data and inference environment can lead to severe performance degradation. Third, many current methods struggle to disentangle speaker identity from linguistic content, resulting in feature entanglement and information leakage, which can compromise speaker anonymity or target similarity.

To address these limitations, we propose **MPFM-VC**, a novel **M**ulti-dimensional **P**erception **F**low **M**atching-based Voice Conversion algorithm. Instead of directly generating speech features or waveforms, MPFM-VC explicitly models the dynamic distribution transfer of speech representations, thereby improving robustness and audio quality under mismatched or noisy conditions. Specifically, MPFM-VC introduces the following core innovations:

- **Multi-dimensional feature flow matching.** We leverage ordinary differential equations (ODEs) to model the dynamic transformation between latent and acoustic spaces, incorporating diverse speech features for adaptive perception and enhancing generation stability.
- **Content perturbation training.** A robustness-aware training scheme is introduced by injecting controlled noise into the content representation, which improves generalization to out-of-distribution data and reduce artifacts such as abnormal plosives during inference.
- **Voiceprint disentanglement based on adversarial learning.** An adversarial training strategy is employed to decouple voiceprint features from linguistic content, reducing cross-feature interference and enhancing target timbre consistency in multi-speaker conversion tasks.

We evaluate MPFM-VC on both speech and singing voice conversion tasks by using a combination of objective metrics and subjective listening tests. Experimental results demonstrate that our model achieves significant improvements in speech quality, robustness, and speaker similarity compared with existing variational and diffusion-based VC methods.

2. Related Work

As shown in Figure 1, recent VC models typically follow a general architecture [1] that comprises a content front-end, an acoustic model, and a vocoder. The content front-end extracts linguistic features from the input speech, which are then transformed into a mel spectrogram by the acoustic model. Finally, the vocoder synthesizes the speech waveform based on the mel spectrogram.

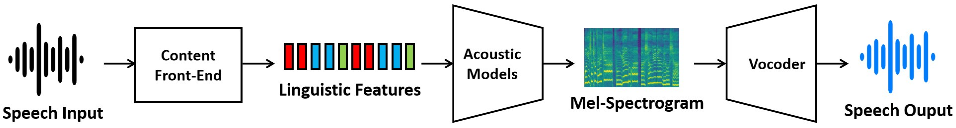


Figure 1. General architecture of VC models.

2.1. End-to-End VC Based on Variational Autoencoder

End-to-end VC approaches integrate all stages of the speech generation pipeline (content front-end, acoustic, and vocoder) into a unified architecture. Recent studies have increasingly adopted the variational autoencoder (VAE) framework as the foundation for such models. A major breakthrough in this direction was the introduction of VITS by SK Telecom [22,23], which unified VAE-based representation learning, stochastic duration modeling, and adversarial training into a single architecture. This integration eliminated the need for explicit alignments and significantly improved synthesis quality, efficiency, and style transfer.

Building upon the VAE framework, FreeVC [24] enhances alignment strategies to achieve higher-quality voice conversion, while Glow-WaveGAN2 [25] improves prosodic control and timbre consistency. DINO-VITS [26] incorporates semi-supervised learning to boost robustness under low-resource conditions. The effectiveness of VITS-based models has been validated against various competitive benchmarks. For instance, in the 2023 Singing Voice Conversion Challenge [27–29], models based on the VITS architecture achieved top rankings in both naturalness and similarity evaluations. Despite their robustness and controllability, current VAE-based systems still face challenges in capturing fine-grained timbral variations and complex semantic representations typical of natural human speech.

2.2. Cascaded VC with Diffusion Models

With the rapid advancement of diffusion models in the field of generative modeling, they are among the mainstream acoustic models within cascaded VC frameworks. DiffWave, introduced by Huawei in 2020 [30], pioneered the use of denoising diffusion probabilistic models (DDPMs) for waveform generation from Gaussian noise. Grad-TTS [31] extended this approach to TTS tasks by applying the diffusion process to spectrum generation, achieving notable improvements in naturalness, prosody control, and temporal variability.

Building upon this foundation, Diff-VC [32] employed diffusion-based denoising to enhance speech naturalness and timbral accuracy, while DDDM-VC [33] further improved clarity and stylistic consistency. ComoSpeech [34] integrated conditional diffusion with a hybrid autoregressive/non-autoregressive decoding strategy, allowing for finer control over emotional expression, prosodic variation, and speaking rate. Models derived from ComoSpeech, including our improved variant, have demonstrated human-comparable performance in subjective evaluations.

Despite these advancements, diffusion-based VC models still face several limitations compared with variational approaches. These include high computational complexity, slower inference speed, and reduced robustness under conditions of distribution shifts.

2.3. Emerging Flow Matching Models

A novel generative paradigm—flow matching—was recently introduced in image synthesis [35] and has gained attention for its efficiency and robustness. Conceptually related to diffusion models, flow matching solves ordinary differential equations (ODEs) to map input noise to output features, enabling faster and more stable generation without iterative sampling. Unlike diffusion processes that rely on stepwise denoising, flow matching enables the linear evolution of features along a predefined trajectory, significantly improving inference speed and reducing computational overhead. Its applicability to multimodal generation—including audio and text—has opened up new opportunities for efficient, high-fidelity speech synthesis.

Recently, researchers in the field of speech generation have also begun to explore the application of flow matching in speech generation [36–39]. For instance, Meta AI proposes Voicebox[36], which is the first speech synthesis system based on non-autoregressive generation combined with a flow matching generation strategy. Voicebox has demonstrated excellent performance, with high robustness, fast inference, cross-lingual generalization ability in speech synthesis tasks, and the flexibility to modify audio segments without re-synthesizing the entire speech in speech-editing tasks.

In this work, we explore the potential of flow matching in voice conversion by integrating it into a cascaded VC architecture. Specifically, we incorporate flow-based dynamic transformation into a mel spectrogram mapping pipeline, aiming to combine the high quality of diffusion models with improved robustness and inference efficiency.

3. Methods

This chapter introduces the proposed voice conversion algorithm, Multi-Dimensional Perception Flow Matching. The overall architecture of the MPFM-VC model is illustrated in Figure 2. In addition to the content features, $I_{content}$, extracted from the input speech, the model also incorporates auxiliary features, such as pitch (I_{pitch}), energy (I_{energy}), and prosody ($I_{prosody}$), to generate the target mel spectrogram, O_{mel}, that corresponds to the speaker's voiceprint, I_{spk}.

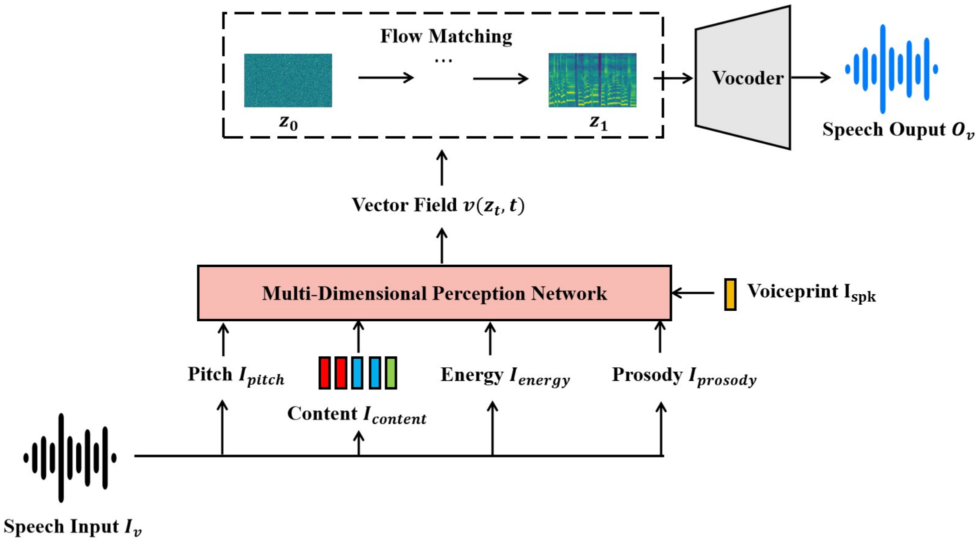


Figure 2. The overall framework of MPFM-VC. Notably, the proposed network, which serves as the acoustic model, does not directly predict the mel spectrogram; instead, it predicts the vector field used in the ODE formulation of flow matching.

In this section, we provide a detailed explanation of the core components of the proposed model, including flow matching for multi-dimensional perception, the multi-dimensional perception network, the content perturbation-based training enhancement method, and the adversarial training mechanism based on the voiceprint.

3.1. Flow matching for multi-dimensional perception

As illustrated in Figure 3, in this study, we adopt a conditionally guided flow matching method based on optimal transport to learn the distribution of mel spectrograms and generate samples from this distribution conditioned on a set of acoustic features. Compared with diffusion probabilistic models [18], the proposed optimal transport-based conditional flow matching approach eliminates the need for reverse processes and complex mathematical derivations. Instead, it generates speech by learning a direct linear mapping between distributions, achieving results comparable to those of diffusion models. This method not only offers better generation performance but also provides a simpler gradient formulation, improved training efficiency, and significantly higher inference speed.

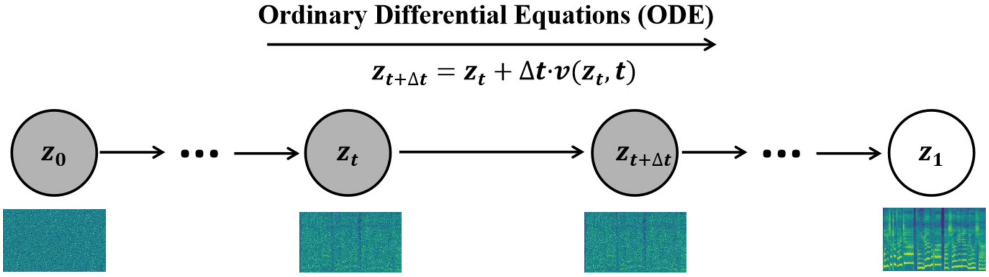

Ordinary Differential Equations (ODE)
$z_{t+\Delta t} = z_t + \Delta t \cdot v(z_t, t)$
z_0
z_t
$z_{t+\Delta t}$
z_1

Figure 3. The process of flow matching.

In this work, data representations are denoted by $z_0 \sim p_0$ and $z_1 \sim p_1$, where p_0 and p_1 represent the prior distribution and the target mel spectrogram distribution, respectively. The subscripts indicate temporal positions, with 0 denoting the starting point and 1 denoting the endpoint. In the flow matching framework, a continuous path of probability densities is constructed from the initial prior distribution $p_0(z \mid z_0) = \mathcal{N}(z \mid 0, 1)$ to the mel spectrogram distribution $p_1(z \mid z_1)$. Notably, the prior distribution in this formulation differs from that of many existing flow matching models in that it is independent of any acoustic-related features. Instead, it is randomly initialized from a standard normal distribution. This acoustically agnostic initialization helps reduce entanglement among the feature representations within the dataset. The entire process can be formally described by the ordinary differential equation (ODE) shown in Equation 1.

$$\frac{dz_t}{dt} = v(z_t, t) \tag{1}$$

where $t \in [0, 1]$ denotes the normalized time and z_t represents the data point at time t. The function $v(z_t, t)$ is a vector field that defines the direction and magnitude of change for each data point in the state space over time. In the flow matching framework, this vector field is parameterized and predicted by a neural network.

Once the predicted vector field $v(z_t, t)$ is obtained, a continuous transformation path from the initial distribution $z_0 \sim p_0$ to the target distribution $z_1 \sim p_1$ can be constructed by solving the corresponding ordinary differential equation (ODE). This ODE can be numerically solved by using the Euler method, as shown in Equation 2.

$$z_{t+\Delta t} = z_t + \Delta t \cdot v(z_t, t) \tag{2}$$

where $\Delta t = 1/N$ denotes the step size, t is the sampled time point, N is the total number of discretization steps, and z_t represents the approximate solution at time t.

Overall, the core idea of flow matching lies in enforcing consistency between the predicted vector field and the ground-truth vector field corresponding to the target mel spectrogram. This ensures that the transformed probability distribution accurately aligns with the desired mel spectrogram distribution. The optimization objective can be formulated as the following loss function:

$$L_{FM}(\theta) = \|v_t(z) - u_t(z)\|^2 \tag{3}$$

where θ denotes the parameters of the neural network, $t \in [0, 1]$, and $z_1 \sim p_1(z)$ represents a sample from the target mel spectrogram distribution. $u_t(z) = z_1 - z_0$ denotes the ground-truth vector field, while $v_t(z)$ represents the predicted vector field to be learned.

3.2. Multi-dimensional perception network

Existing flow matching models typically adopt a U-Net architecture to predict the vector field [40]. In this work, we propose a novel multi-dimensional feature perception network, as illustrated in Figure 4, which enhances the model's ability to handle diverse acoustic conditions. Prior to being fed into the proposed feature perception blocks, all inputs undergo an encoding process as follows: First, a Transformer[17]-based sinusoidal positional encoding layer is introduced to generate time embeddings E_t. Second, speaker identity features I_{spk} are projected onto a latent space via a linear layer to produce speaker embeddings E_{spk}. For the complex content representation $I_{content}$, a Transformer encoder is employed to obtain content embedding $E_{content}$. Simultaneously, pitch sequences I_{pitch} are processed through a Transformer embedding layer to yield pitch embeddings E_{pitch}. Additionally, at each time step t in the flow matching process, the intermediate sample z_t is computed by using linear interpolation: $z_t = z_0 \cdot t + z_1 \cdot (1 - t)$. Finally, the above multimodal embeddings are concatenated along the feature dimension and are used as inputs to the feature perception block.

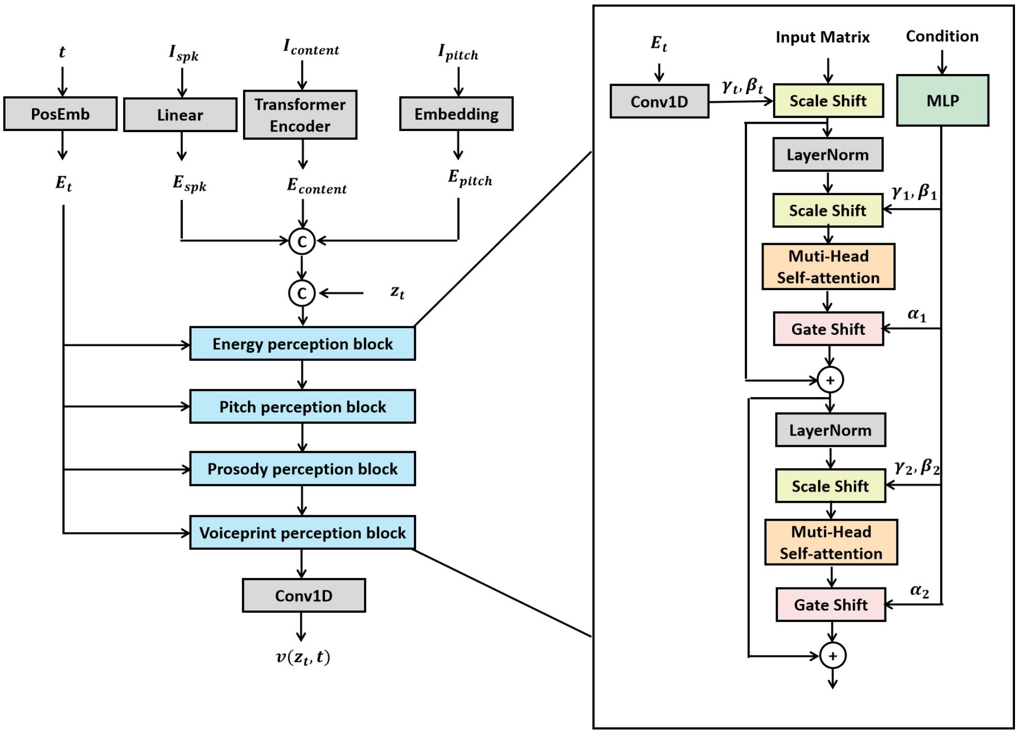


Figure 4. The architecture of the multi-dimensional feature perception network.

The multi-dimensional perception network proposed in this study adopts a modular architecture composed of multiple feature-awareness blocks. While all blocks share the same structural design, they utilize distinct condition matrices to capture diverse contextual information. Each block integrates conditional information through two complementary mechanisms:

(1) **Scale Shift**, defined in Equation 4, modulates the feature distribution by applying learnable scaling and bias transformation.

(2) **Gated Shift**, formulated in Equation 5, employs a gated unit to dynamically regulate the flow of information through adaptive feature reweighting.

The computations of these mechanisms are given by

$$ScaleShift(x) = x * \gamma + \beta \tag{4}$$

$$GateShift(x) = x(1 + \alpha) + \alpha \tag{5}$$

where α, γ, and β are learnable parameters of the network and x denotes the input matrix.

Specifically, the embedding of conditional information is carried out through a hierarchical processing mechanism. First, the time encoding is transformed by a one-dimensional convolutional layer to produce two learnable parameters, (γ_t and β_t), which are used for time-specific feature modulation. Next, the condition matrix is passed through a multi-layer perceptron (MLP) and mapped to six adaptive parameters ($\alpha_1, \gamma_1, \beta_1, \alpha_2, \gamma_2$, and β_2), enabling conditional feature encoding. The network then executes the following steps sequentially: (1) the application of layer normalization to the input feature matrix followed by conditional feature modulation; (2) cross-dimensional feature interaction based on a multi-head self-attention mechanism; (3) the application of a gated control mechanism to regulate the flow of information. To ensure training stability, a residual connection is introduced at the end of the block. This not only preserves the original feature information but also effectively mitigates issues related to gradient instability.

In the proposed multi-dimensional perception network, the condition matrix is used to embed auxiliary information, including energy (I_{energy}), prosody ($I_{prosody}$), pitch (I_{pitch}), and speaker voiceprint (I_{spk}). These conditioning features are incorporated into different feature-awareness blocks, each responsible for modeling a specific type of information. Specifically, the first feature-awareness block utilizes the energy embedding, E_{energy}, to inject energy-related information; the second block employs the pitch embedding, E_{pitch}, to capture pitch dynamics; the third block uses the prosody embedding, $E_{prosody}$, for prosodic modeling; and the final block integrates speaker characteristics by using the speaker embedding, E_{spk}. These encoded features are fused through a one-dimensional convolutional layer and are ultimately used to predict the vector field, $v(z_t, t)$. The core design principle behind this structure is that the closer to the output layer a block is, the stronger its sensitivity to conditional information is. Therefore, in acoustic modeling tasks, the speaker's timbre should be assigned greater importance, while energy-related information can be considered relatively less influential. Nonetheless, each feature-awareness block is capable of adaptively learning and adjusting its sensitivity to various types of conditioning through dynamic parameterization, enabling the model to flexibly control the contribution of each feature and optimize feature fusion strategies based on task-specific requirements.

In conventional flow matching models, the vector field, $v_t(z)$, is typically trained uniformly across all time steps. However, in practice, predictions at intermediate time steps tend to be more challenging. Specifically, when t approaches 0, the optimal prediction tends to align with the mean of the target distribution (p_1), while for t close to 1, it tends to align with the mean of the prior distribution (p_0). In contrast, predictions around the midpoint ($t \approx 0.5$) are often more ambiguous and unstable due to increased distributional uncertainty. To more accurately learn the ground-truth vector field, $u_t(z) = z_1 - z_0$, we introduce a log-normal weighting scheme in the time dimension to reweight the loss function within the optimal transport-based flow matching framework. This adjustment emphasizes the more difficult training samples at intermediate time steps, thereby improving the learning of vector dynamics during these transitions. The weighted loss function is defined as follows:

$$L_{FMLog}(\theta) = \frac{1}{\sqrt{2\pi}} \frac{1}{t(1-t)} exp\left(-\left(\log \frac{t}{1-t}\right)^2\right) \|v(z(t), t) - (z_1 - z_0)\|^2 \tag{6}$$

The log-normal distribution assigns lower loss weights to intermediate time steps, making them easier to optimize during training. In contrast, higher loss weights are assigned to time steps near 0 and 1, encouraging the model to converge more rapidly to optimal solutions in those regions.

3.3. Content perturbation-based training enhancement method

Due to distributional discrepancies between the training data and the inference environment, the model may produce degraded audio quality during inference, such as unnatural artifacts or plosive distortions. To improve generalization and robustness, we propose a content perturbation-based training enhancement method. This method is designed to enhance the model's contextual generalization capability and improve its stability in real-world deployment scenarios.

As illustrated in Figure 5, random perturbations are applied to the input content representations during the training phase to enhance model robustness. Given an input content representation vector $I_{content}$, a certain proportion of its dimensions are randomly selected and masked. The selected dimensions are replaced with blank tokens, and the perturbation strategy is defined as follows:

$$I'_{content} = M \odot I_{content} + (1 - M) \odot I_{blank} \tag{7}$$

where $M \in 0, 1^d$ denotes a binary masking vector, where a subset of elements is randomly set to 0 according to a predefined masking ratio, while the remaining elements are set to 1. I_{blank} refers to the blank token used for replacement during perturbation, which is generated based on a roulette-wheel sampling strategy. Specifically, the value distribution of the content representation is first estimated from the training data by computing the frequency of each value within the embedding space. Then, values are randomly sampled in proportion to their observed frequencies to construct the blank representation used for masking.

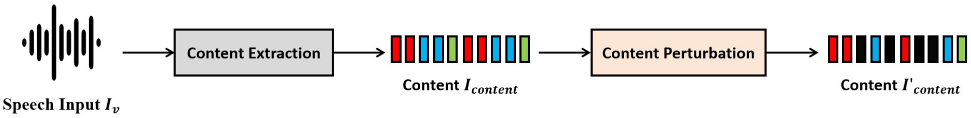


Figure 5. Content perturbation-based training enhancement method.

Through this training strategy, the model is expected to generate high-quality speech even when partial content representations are missing. This enables the model to maintain strong robustness and generalization during inference, particularly in scenarios involving incomplete feature inputs or noise-corrupted conditions, thereby ensuring stable and natural speech synthesis.

3.4. Adversarial training mechanism based on voiceprints

In VC models, the primary objective of the content encoder is to extract speaker-independent content representations. However, during training, since the target and reference speakers are often the same, the content encoder inevitably captures speaker-specific information. This leakage results in synthesized speech that still carries residual timbral characteristics of the reference speaker, thereby degrading the naturalness and quality of the generated audio. To address this issue, we introduce an adversarial training strategy that employs a gradient reversal layer (GRL) to explicitly disentangle speaker identity information from the content representation, facilitating the generation of speaker-independent content features.

As illustrated in Figure 6, the overall adversarial training framework consists of three main components: the content encoder ($E_{content}$) from the flow matching module, a gradient reversal layer (GRL), and a speaker classifier ($Classifier_{spk}$). Specifically, within the acoustic model, the input content representation, $I_{content}$, is first processed by the content encoder, $Encoder_{content}$, to generate a speaker-independent content embedding:

$$E_{content} = Encoder_{content}(I_{content}) \tag{8}$$

To disentangle speaker-related information from the content encoder, we design a speaker classifier composed of three linear layers with ReLU activation functions. This classifier aims to predict the speaker identity, P_{spk}, from the content embedding, $E_{content}$, and guides the training process by comparing the prediction with the ground-truth speaker embedding, I_{spk}.

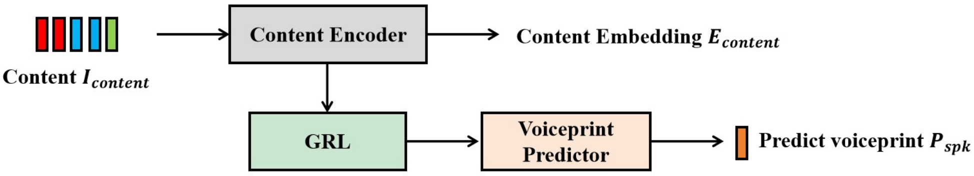


Figure 6. Adversarial training mechanism based on voiceprint.

To enforce feature disentanglement, a gradient reversal layer (GRL) is inserted before the speaker classifier. During forward propagation, the GRL passes the content representation unchanged. However, during backpropagation, it reverses the gradient direction and scales it by a tunable adversarial coefficient λ:

$$\frac{\partial L_{spk}}{\partial I_{content}} \leftarrow -\lambda \cdot \frac{\partial L_{spk}}{\partial I_{content}} \tag{9}$$

where λ denotes the adversarial coefficient, which controls the strength of gradient reversal.

The speaker classifier, $Classifierspk$, is optimized based on this mechanism to accurately predict the speaker identity by minimizing the loss, L_{spk}, while the content encoder, $Encoder_{content}$, is trained adversarially to maximize this loss—effectively preventing the classifier from extracting speaker-specific cues. This adversarial game facilitates the disentanglement of speaker information from content embeddings. The final training objective is defined as

$$\min_{Encoder_{content}} \max_{Classifier_{spk}} L_{spk} = \frac{P_{spk} \cdot I_{spk}}{|P_{spk}| \times |I_{spk}|} \tag{10}$$

3.5. The training and inference processes of MPFM-VC

Algorithm 1 The inference process of MPFM-VC.

```
Require: $MPFM(\theta)$; $Vocoder(\theta)$; Input speech $I_v$; voiceprint $I_{spk}$ and the maximum number of sampling steps $N_{max}$.
Initial time step $t = 0$, time step $\Delta t = 1/N_{max}$
Extracted from the original speech $(I_{content}, I_{pitch}, I_{energy}, I_{prosody})$
Samples $z_0$ from a random normal distribution $N(0,1)$
for $i = 1, 2, \ldots, N_{max}$ do
    $z_{t+\Delta t} = z_t + \Delta t \cdot MPFM(z_t, t, I_{content}, I_{spk}, I_{pitch}, I_{energy}, I_{prosody})$
    $t = t + \Delta t$
end for
$O_{mel} = z_1$
$O_v = Vocoder(O_{mel}, I_{pitch})$
return $O_v$
```

Algorithm 2 The training process of MPFM.

```
Require: $MPFM(\theta)$; Dataset $D_{train} = \{(I_{content}, I_{spk}, I_{pitch}, I_{energy}, I_{prosody}, T_{mel})\}_{m=1}^{M}$; Number of training rounds $N_{iter}$; Learning rate $\eta$;
for $i = 1, 2, \ldots, N_{iter}$ do
    From $D_{train}$ sample $(I_{content}, I_{spk}, I_{pitch}, I_{energy}, I_{prosody}, T_{mel})$
    Randomly sample $t$ from $[0, 1]$
    Samples $z_0$ from a random normal distribution $N(0,1)$
    Calculate $z_t = z_0 + (z_1 - z_0) * t$
    Content perturbation training augmentation method:
      $I'_{content} = M \odot I_{content} + (1 - M) \odot I_{blank}$
    Forward propagation: $v(z_t, t), P_{spk} = MPFM(z_t, t, I'_{content}, I_{spk}, I_{pitch}, I_{energy}, I_{prosody})$
    Calculates the loss function: $L_{total} = L_{FMLog} - L_{spk}$
    Backward propagation: Calculate the gradient $\nabla_\theta L_{total}$
    Take gradient descent step on $\theta \leftarrow \theta - \eta \cdot \nabla_\theta L_{total}$
end for
return $MPFM(\theta)$
```

4. Results

4.1. Dataset

We conducted experiments on a Mandarin multi-speaker speech dataset, AISHELL-3[41], and a Mandarin multi-singer singing dataset, M4Singer[42], to evaluate the effectiveness of the proposed method in both speech and singing voice conversion tasks.

(1) AISHELL-3

AISHELL-3 is a high-quality multi-speaker Mandarin speech synthesis dataset released by AISHELL Foundation. It contains 218 speakers (where the male-to-female ratio is balanced) and more than 85000 speech sentences, the total duration is about 85 hours, and the sampling rate is 44.1kHz. A 16-bit, professional recording environment made clear sound quality possible. In order to construct the speech test set, in this study, we randomly selected 10 of the speakers to be excluded from training, and from each, we randomly selected 10 speech samples, totaling 100 speech samples.

(2) M4Singer

M4Singer is a large-scale Chinese singing voice dataset with multiple styles and multiple singers released by Tsinghua University. The dataset contains 30 professional singers (where the male-to-female ratio is balanced) and a total of 16000 singing sentences, the total duration is 18.8 hours, and the sampling rate is 44.1kHz. It was generated in a 16-bit, clean, and noise-free recording environment and covers pop, folk, rock, and other styles. Similar to the above, 10 singers were randomly selected to be excluded from training, and 10 singing samples were selected for each singer, for a singing test set totaling 100 samples.

4.2. Data processing

In this study, multiple key features were extracted from the original speech data, including content representation ($I_{content}$), speaker voiceprint (I_{spk}), pitch (I_{pitch}), energy (I_{energy}), prosody ($I_{prosody}$), and the target mel spectrogram. The content representation, $I_{content}$, was extracted by using a pre-trained automatic speech recognition model, SenseVoice[43], which provides high-precision linguistic features. Speaker identity features I_{spk} were obtained by using the pre-trained speaker verification model Camplus[44], which captures speaker-dependent characteristics. Pitch information I_{pitch} was extracted by using the pre-trained neural pitch estimation model RMVPE[45], which directly derives pitch features from raw audio, ensuring high accuracy and robustness. Energy I_{energy} was calculated as the root mean square (RMS) energy of each frame in the speech signal. Prosodic features $I_{prosody}$ were extracted by using the neural prosody encoder HuBERT-Soft[46], which captures rhythm, intonation, and other prosodic cues in speech. The target mel spectrogram was computed with a standard signal processing pipeline consisting of pre-emphasis, framing, windowing, Short-Time Fourier Transform (STFT), power spectrum calculation, and mel filterbank projection. The configuration details are as follows: a sampling rate of 32,000 Hz, a pre-emphasis coefficient of 0.97, a frame length of 1024, a frame shift of 320, the Hann window function, and 100 mel filterbank channels.

4.3. Model parameters

Regarding feature input parameters, the speaker voiceprint (I_{spk}) had a dimension of 192 and was first projected into a 100-dimensional speaker embedding (E_{spk}) with a linear layer. The content representation ($I_{content}$) is a one-dimensional sequence of length T with a vocabulary size of 4096. It was embedded into a 512-channel matrix with an embedding layer and then fed into a Transformer-based encoder consisting of six blocks. Each block contained eight attention heads and a feed-forward layer with a hidden dimension of 2048. The output was content embedding $E_{content}$. The pitch feature (I_{pitch}) was first mapped to a discrete sequence with a vocabulary size of 256, which was then transformed into a 512-channel matrix by using an embedding layer. The energy (I_{energy}) and prosody ($I_{prosody}$) features were both projected into 100-dimensional embeddings (E_{energy} and $E_{prosody}$, respectively) by using linear layers.

In the multi-dimensional perception network, the time encoding was processed by a one-dimensional convolutional layer with an output size of 512×2 to generate two modulation parameters, γ_t and β_t. The multi-head self-attention module used four attention heads with a hidden dimension of 400. The conditional embedding network consisted of a multi-layer perceptron with two linear layers, where the hidden layer had a dimensionality of 400, and the output layer had a size of 400×6, producing six conditional parameters: $\alpha_1, \gamma_1, \beta_1, \alpha_2, \gamma_2$, and β_2. Finally, the entire multi-dimensional feature perception network output a 100-dimensional predicted vector field ($v(z_t, t)$) via a one-dimensional convolutional layer.

4.4. Training Setup

The experiments were conducted on both speech and singing datasets, with training and testing performed separately for speech conversion and singing voice conversion tasks. The model was trained for 100 epochs by using the Adam optimizer, until full convergence was achieved. A dynamic batch size strategy was adopted, where the batch size was determined based on the total frame length of the content representation ($I_{content}$), with a maximum limit of 10,000 frames per batch. The learning rate was set by using a warm-up strategy, with an initial learning rate $lr_i = 0.001$ and a warm-up step count of 2,500. The learning rate at step t is computed as follows:

$$lr = lr_i * warmup^{0.5} * min(step^{-0.5}, step * warmup^{-1.5}) \tag{11}$$

4.5. Baseline models and evaluation metrics

(1) Baseline models

1) Free-VC[24] (2022, ICASSP) is a voice conversion model that adopts the variational autoencoder architecture of VITS for high-quality waveform reconstruction; it is widely used in voice conversion tasks due to its efficient feature modeling capabilities.

2) Diff-VC[32] (2022, ICLR) is a diffusion model-based voice conversion method which can generate high-quality converted speech based on noise reconstruction; it is the representative diffusion model for VC tasks.

3) DDDM-VC[33] (2024, AAAI) is a newly proposed feature decoupling speech conversion method based on a diffusion model which improves the quality of converted speech and speaker consistency while maintaining the consistency of speech features.

(2) Evaluation index

1) Mean Opinion Score (MOS): The naturalness of the synthesized speech was evaluated by 10 students with good Mandarin skills and sound sense as the audience.

2) Mel Cepstral Distortion (MCD): It measures the spectral distance between the converted speech and the target speech, where a lower value indicates higher-quality conversion.

3) Word Error Rate (WER): The intelligibility of the converted speech was evaluated by using automatic speech recognition, where a lower WER indicates higher speech intelligibility.

4) Similarity Mean Opinion Score (SMOS): Listeners (10 students with good Mandarin skills and sound sense) scored the timbre similarity of the synthesized speech for us to measure the subjective similarity of the timbre after speech conversion.

5) Speaker Embedding Cosine Similarity (SECS): The cosine similarity between the original speech and the converted speech was calculated based on the speaker coding, which is used to objectively measure the degree of timbre preservation. The higher the value is, the closer the converted speech is to the target timbre.

4.6. Experimental results

(1) Quality evaluation of voice conversion

The primary goals of the voice conversion (VC) task are to transform the speaker's timbre while preserving the original linguistic content and maximize the naturalness and intelligibility of the generated speech. To this end, we evaluated the proposed method on the speech test set by using both subjective and objective metrics. Subjective evaluation was conducted based on Mean Opinion Score (MOS) tests, while objective performance was quantified by using Mel Cepstral Distortion (MCD) and the Word Error Rate (WER). This combination provides a comprehensive assessment of the effectiveness of different voice conversion approaches.

As shown in Table 1, the proposed algorithm, MPFM-VC, demonstrates superior performance in the voice conversion task, achieving the highest scores in terms of naturalness, audio fidelity, and speech clarity. Compared with existing methods, MPFM-VC exhibits stronger stability across multiple evaluation metrics, indicating its ability to maintain high-quality synthesis under varying data conditions. Specifically, MPFM-VC achieves an 11.57% improvement in the MOS over Free-VC, showing significant advantages in speech continuity and prosody control and effectively avoiding the distortion issues commonly observed in traditional end-to-end VITS-based frameworks. In comparison with diffusion-based models such as Diff-VC and DDDM-VC, MPFM-VC achieves the lowest MCD (6.23) and the lowest WER (4.23%), which suggests that it better preserves the semantic content of the target speaker during conversion, thereby enhancing the intelligibility and clarity of the generated speech. These results highlight that the integration of multi-dimensional feature perception modeling and content perturbation-based training augmentation significantly improves the model's ability to adapt to various speech features. Consequently, MPFM-VC delivers consistent and high-quality voice synthesis across different speakers and speaking contexts.

Table 1. Results of voice conversion quality evaluation.

Evaluated Models	MOS (↑)	MCD (↓)	WER (↓)
GT	4.32 ± 0.05	-	1.79%
Free-VC	3.63 ± 0.04	7.11	13.28%
Diff-VC	3.87 ± 0.04	6.79	11.27%
DDDM-VC	3.91 ± 0.05	6.44	5.24%
MPFM-VC (ours)	$\mathbf{4.05 \pm 0.05}$	**6.23**	**4.23%**

(2) Timbre similarity evaluation of voice conversion

The goal of speaker similarity evaluation is to assess a voice conversion method's ability to preserve the timbre consistency of the target speaker. In this study, we adopt two metrics for analysis: Similarity MOS (SMOS) and Speaker Embedding Cosine Similarity (SECS).

As shown in Table 2 and Figure 7, MPFM-VC demonstrates strong speaker consistency in the speaker similarity evaluation task, achieving the highest SMOS (3.83) and SECS (0.84) scores. This indicates that MPFM-VC is more effective in preserving the timbral characteristics of the target speaker during conversion. Compared with Free-VC, MPFM-VC enhances the model's adaptability to target speaker embeddings through multi-dimensional feature perception modeling, thereby improving post-conversion timbre similarity. Although Diff-VC benefits from diffusion-based generation, which improves overall audio quality to some extent, it fails to sufficiently disentangle speaker identity features, resulting in residual characteristics from the source speaker in the converted speech. While DDDM-VC introduces feature disentanglement mechanisms that improve speaker similarity, it still falls short compared with MPFM-VC. These findings suggest that the combination of flow matching modeling and adversarial training on speaker embeddings in MPFM-VC effectively suppresses unwanted speaker information leakage during conversion. As a result, the synthesized speech is perceptually closer to the target speaker's voice while maintaining naturalness and improving controllability and stability in voice conversion tasks.

Table 2. Results of voice conversion timbre similarity evaluation.

Evaluated Models	SMOS (↑)	SECS (↑)
Free-VC	3.28 ± 0.05	0.72
Diff-VC	3.51 ± 0.03	0.79
DDDM-VC	3.67 ± 0.04	0.81
MPFM-VC (ours)	$\mathbf{3.83 \pm 0.04}$	**0.84**

doi:10.20944/preprints202504.1428.v2

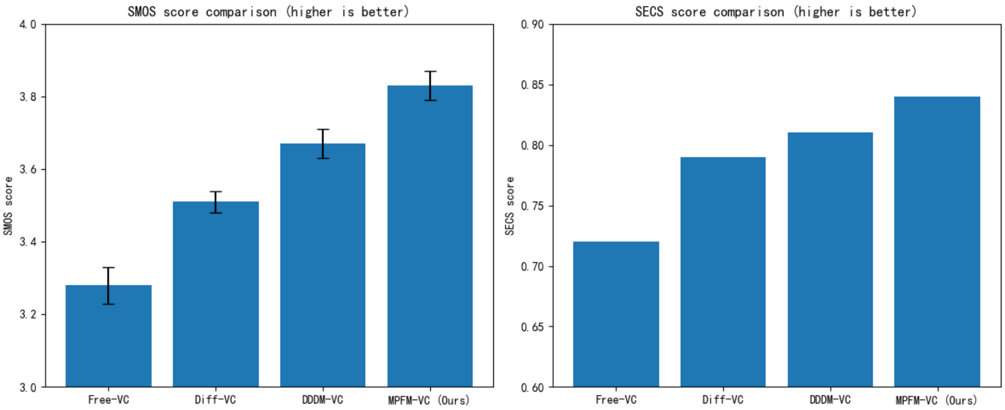


Figure 7. Results of voice conversion timbre similarity evaluation.

(3) Quality evaluation of singing voice conversion

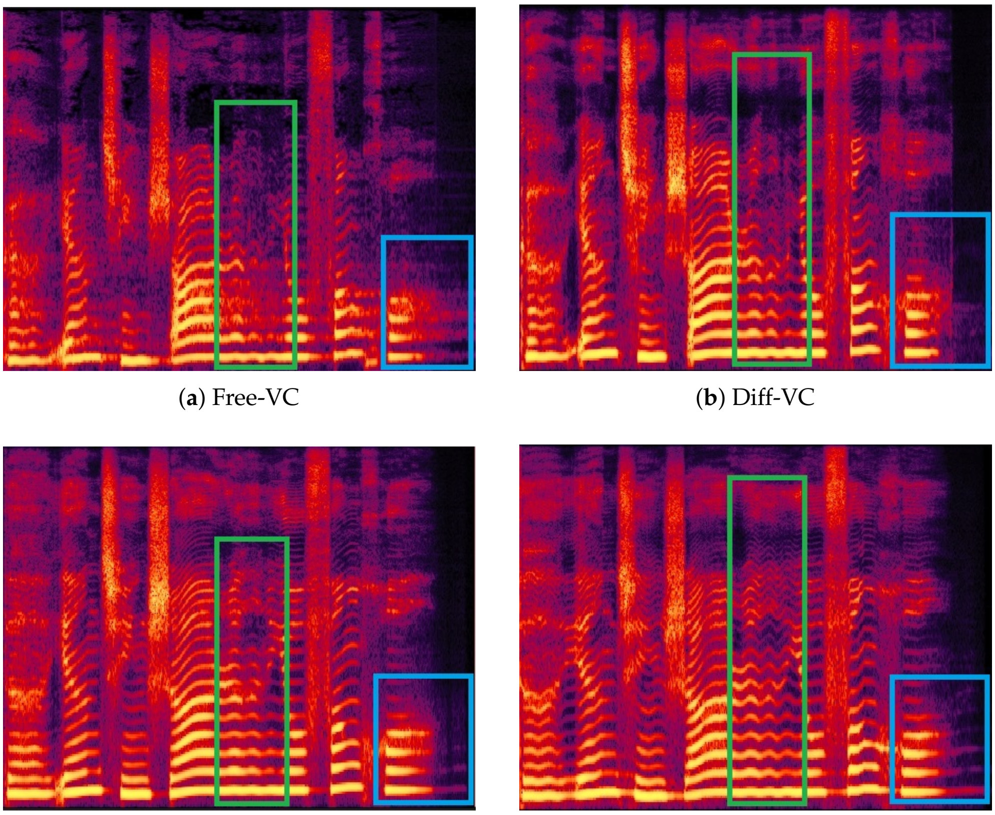
In the context of voice conversion, singing voice conversion is generally more challenging than standard speech conversion due to its inherently richer pitch variations, timbre stability, and prosodic complexity. We adopted the same set of evaluation metrics used in speech conversion to comprehensively evaluate the performance of different models on the singing voice conversion task, including the subjective Mean Opinion Score (MOS) and objective indicators such as Mel Cepstral Distortion (MCD) and the Word Error Rate (WER).

As shown in Table 3, MPFM-VC also demonstrates outstanding performance on the singing voice conversion task, achieving a MOS of 4.12, an MCD of 6.32, and a WER of 4.86%.Through multi-dimensional feature perception modeling, MPFM-VC effectively adapts to melodic variations and pitch fluctuations inherent in singing voices, resulting in converted outputs that are more natural and fluent while maintaining high levels of audio quality and clarity. Compared with Free-VC, MPFM-VC further improves the naturalness of the generated singing voices by leveraging flow matching, which enhances the modeling of dynamic acoustic features during conversion. In contrast to diffusion-based methods such as Diff-VC and DDDM-VC, MPFM-VC avoids the timbre over-smoothing often introduced by diffusion models, which can lead to the loss of fine-grained acoustic details. As a result, the synthesized singing voices generated by MPFM-VC exhibit greater depth, expressiveness, and structural richness.

Table 3. Results of singing voice conversion quality evaluation.

Evaluated Models	MOS (↑)	MCD (↓)	WER (↓)
GT	4.39 ± 0.04	-	2.12%
Free-VC	3.52 ± 0.04	7.09	18.28%
Diff-VC	3.73 ± 0.05	6.99	13.27%
DDDM-VC	3.81 ± 0.04	6.64	7.24%
MPFM-VC (ours)	$\mathbf{4.12 \pm 0.06}$	**6.32**	**4.86%**

Additionally, we randomly selected a singing voice segment for spectrogram comparison, as shown in Figure 8. The proposed MDFM-VC produces the clearest and most well-defined spectrogram, benefiting from the additional feature inputs of the multi-dimensional perception network, which allow the model to reconstruct finer acoustic details such as vibrato and tail articulations more accurately. In contrast, although DDDM-VC and Diff-VC are also capable of generating spectrograms with strong intensity and clarity, their outputs tend to suffer from over-smoothing, which results in the loss of important contextual and prosodic information—diminishing the expressive detail in the converted singing voice.



(**a**) Free-VC

(**b**) Diff-VC

(**c**) DDDM-VC

(**d**) MDFM-VC

Figure 8. Spectrogram comparison for the same voice input.

(4) Timbre similarity evaluation of singing voice conversion

Timbre similarity in singing voices is a critical metric for evaluating a model's ability to preserve the target singer's vocal identity during conversion. Compared with normal speech, singing involves more complex pitch variations, formant structures, prosodic patterns, and timbre continuity, which pose additional challenges for accurate speaker similarity modeling. In this study, we performed a comprehensive evaluation by using both subjective SMOS (Similarity MOS) and objective SECS (Speaker Embedding Cosine Similarity) to assess the effectiveness of different methods in capturing and preserving timbral consistency in singing voice conversion.

As shown in Table 4 and Figure 9, MPFM-VC achieves superior performance in singing voice timbre similarity evaluation. It outperforms traditional methods in both SMOS (3.79) and SECS (0.81), indicating its ability to more accurately preserve the target singer's timbral identity. In singing voice conversion, Free-VC and Diff-VC fail to sufficiently disentangle content and speaker representations, leading to perceptible timbre distortion and poor alignment with the target voice. Although the diffusion-based DDDM-VC model partially solves this issue, it still suffers from timbre over-smoothing, resulting in synthesized singing voices that lack distinctiveness and individuality. In contrast, MPFM-VC incorporates an adversarial speaker disentanglement strategy, which effectively suppresses residual source speaker information and ensures that the converted singing voice more closely resembles the target singer's timbre. Additionally, MPFM-VC enables the fine-grained modeling of timbral variation with a multi-dimensional feature-aware flow matching mechanism, leading to improved timbre stability and consistency throughout the conversion process.

Table 4. Results of singing voice conversion timbre similarity evaluation.

Evaluated Models	SMOS (↑)	SECS (↑)
Free-VC	3.32 ± 0.05	0.67
Diff-VC	3.42 ± 0.04	0.72
DDDM-VC	3.57 ± 0.05	0.77
MPFM-VC (ours)	**3.79 ± 0.03**	**0.81**

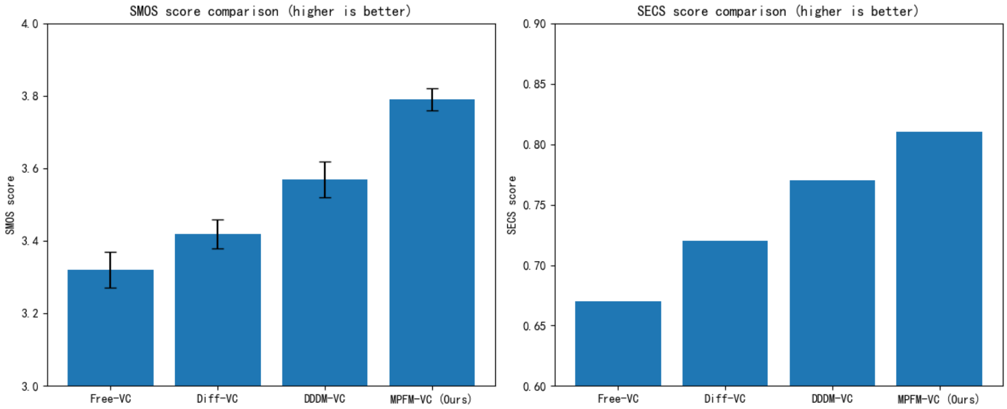


Figure 9. Results of singing voice conversion timbre similarity evaluation.

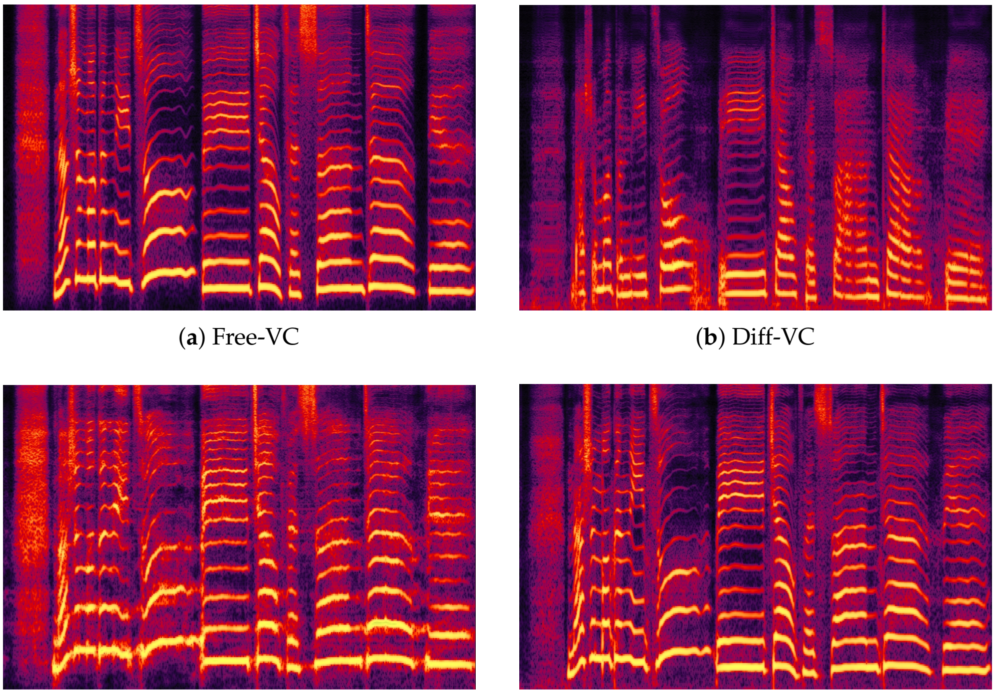
(5) Robustness Evaluation under Low-Quality Conditions

In real-world applications, voice conversion systems must exhibit robustness to low-quality input data in order to maintain reliable performance under adverse conditions such as background noise, limited recording hardware, or unclear articulation by the speaker. To assess this capability, we additionally collected a set of 30 low-quality speech samples that incorporate common noise-related challenges, including mumbling, background reverberation, ambient noise, signal clipping, and low-bitrate encoding.

As shown in Table 5 and the spectrograms in Figure 10, the proposed algorithm, MPFM-VC, demonstrates strong performance even under low-quality speech conditions. The generated spectrograms remain sharp and well defined, indicating high-fidelity synthesis, whereas other voice conversion systems exhibit significantly degraded robustness—resulting in reduced naturalness and timbre consistency in the converted outputs. In particular, diffusion-based models such as Diff-VC and DDDM-VC suffer from substantial performance degradation in noisy environments, with spectrograms appearing blurry and incomplete. This suggests that diffusion models have limited stability under extreme data conditions and are less effective in handling perturbations introduced by low-quality inputs. Moreover, Diff-VC performs the worst in both the MCD and WER metrics, indicating a large mismatch between its generated mel spectrograms and the target speech, as well as a severe decline in speech intelligibility. These results reveal that Diff-VC is highly sensitive to input noise, making it less suitable for real-world applications where input quality cannot be guaranteed.

Table 5. Results of voice conversion under low-quality conditions.

Evaluated Models	MOS (↑)	MCD (↓)	WER (↓)	SMOS (↑)	SECS (↑)
GT	3.55 ± 0.03	-	21.21%	3.73 ± 0.06	0.82
Free-VC	3.07 ± 0.03	7.24	25.82%	3.21 ± 0.04	0.62
Diff-VC	2.49 ± 0.06	8.85	35.20%	2.21 ± 0.05	0.52
DDDM-VC	2.87 ± 0.06	8.42	32.61%	2.89 ± 0.05	0.59
MPFM-VC (ours)	**3.28 ± 0.03**	**6.89**	**15.37%**	**3.53 ± 0.05**	**0.73**



(**a**) Free-VC (**b**) Diff-VC

(**c**) DDDM-VC (**d**) MPFM-VC

Figure 10. Spectrograms under low-quality conditions.

It is worth noting that although Free-VC showed relatively weaker performance in previous experiments, it still outperforms diffusion-based architectures under low-quality speech conditions. Its generated spectrograms appear only slightly blurred, indicating that the end-to-end variational autoencoder (VAE)-based modeling approach offers a certain degree of robustness to noise. However, its SECS score remains significantly lower than that of MPFM-VC, suggesting persistent inaccuracies in timbre matching.

In contrast, MPFM-VC consistently maintains superior speech quality and speaker consistency even under low-quality input conditions. It achieves the best performance across all evaluation metrics—including MOS, MCD, WER, and SMOS—and its spectrograms remain sharp and vibrant. This advantage can be attributed to the multi-dimensional feature-aware flow matching mechanism, which enables the fine-grained modeling of speech features under varying noise conditions. Additionally, the content perturbation-based training augmentation strategy allows the model to adapt to incomplete or degraded inputs during training, resulting in greater robustness during inference. Furthermore, the adversarial training on speaker embeddings enhances timbre preservation under noisy conditions, allowing MPFM-VC to significantly outperform other methods in SECS and more accurately retain the target speaker’s timbral characteristics.

4.7. Ablation experiments

(1) Content perturbation-based training enhancement method

In the voice conversion task, the content perturbation-based training augmentation strategy is designed to improve model generalization by introducing controlled perturbations to content representations during training. This approach aims to reduce inference-time artifacts such as unexpected noise and improve the overall stability of the converted speech. To validate the effectiveness of this method, we conducted an ablation study by removing the content perturbation mechanism and observing its impact on speech conversion performance. A test set consisting of 100 out-of-distribution audio samples with varying durations was used to simulate real-world scenarios under complex conditions. The following evaluation metrics were determined: MOS, MCD, WER, SMOS, SECS, and the frequency of plosive artifacts per 10 seconds in converted speech (Badcase).

As shown in Table 6, the content perturbation-based training augmentation strategy plays a crucial role in improving the stability and robustness of the voice conversion model. Specifically, after removing this module, the Badcase rate increases significantly—from 0.39 to 1.52 occurrences per 10 seconds—indicating a higher frequency of artifacts such as plosive noise, interruptions, or other unexpected distortions under complex, real-world conditions. In addition, both MCD and WER show slight increases, suggesting a decline in the acoustic fidelity and intelligibility of the converted speech.

Table 6. Results of ablation experiments for content perturbation-based training enhancement method.

Evaluated Metrics	MPFM-VC (w/ content perturbation)	MPFM-VC (w/o content perturbation)
Badcase (times/10s)	0.39	1.52
MOS	3.85 ± 0.04	3.92 ± 0.05
MCD	6.32	6.47
WER	3.87%	3.92%
SMOS	3.79 ± 0.06	3.72 ± 0.04
SECS	0.81	0.82

Interestingly, a minor improvement is observed in the MOS, which may be attributed to the model's tendency to overfit the training distribution when not exposed to adversarial or perturbed inputs. As a result, the model performs better on in-domain evaluation sets but suffers from reduced generalization and greater output variability when tested on more challenging, out-of-distribution audio samples with varying durations. It is also worth noting that the SMOS and SECS remain largely unchanged, implying that content perturbation primarily contributes to improving speech stability, rather than influencing timbre consistency.

In summary, the proposed content perturbation strategy effectively reduces unexpected artifacts and enhances the stability and generalization capability of the voice conversion system. These findings confirm that incorporating this method is critical to maintaining high speech quality under diverse and noisy-input conditions and thus holds significant practical value in real-world deployment scenarios.

(2) Adversarial training mechanism based on voiceprint

In the voice conversion task, the strategy of adversarial training on speaker embeddings is designed to enhance the timbre similarity to the target speaker while suppressing residual speaker identity information from the source speaker. This ensures that the converted speech better matches the target speaker's voice without compromising the overall speech quality. To evaluate the effectiveness of this strategy, we conducted an ablation study by removing the adversarial training module and comparing its impact on timbre similarity and overall speech quality in the converted outputs.

As shown in Table 7, the strategy of adversarial training on speaker embeddings plays a critical role in improving timbre matching in the voice conversion task. Specifically, after removing this module, both the SMOS and SECS drop significantly—the SMOS decreases from 3.83 to 3.62, and SECS drops from 0.84 to 0.73. This indicates a notable decline in target timbre consistency, as the

Table 7. Ablation experimental results of adversarial training method for voiceprint features.

Evaluated Metrics	MPFM-VC (w/ voiceprint adversarial)	MPFM-VC (w/o voiceprint adversarial)
MOS	4.05 ± 0.03	3.82 ± 0.04
MCD	6.23	6.37
WER	4.23%	3.72%
SMOS	3.83 ± 0.05	3.62 ± 0.06
SECS	0.84	0.73

converted speech becomes more susceptible to residual characteristics from the source speaker, leading to suboptimal conversion performance.

On the other hand, the MOS and MCD remain largely unchanged, suggesting that adversarial training primarily enhances timbre similarity without significantly affecting overall speech quality. Interestingly, the WER shows a slight improvement, implying that removing the adversarial mechanism may enhance intelligibility and clarity. However, this improvement likely comes at the cost of timbre fidelity; in other words, while the output speech may sound clearer, it deviates more from the target speaker's vocal identity, resulting in less precise conversion.

Overall, the speaker adversarial training strategy ensures accurate timbre alignment by suppressing residual speaker identity features from the source, making the converted speech more consistent with the desired voice. Although some quality metrics slightly improve when the module is removed, this is primarily due to the model reverting to generic or averaged timbre features, rather than capturing the distinct timbral traits of the target speaker. Therefore, in practical applications, adversarial training remains essential to achieving high-quality voice conversion—ensuring that the generated speech is not only intelligible but also accurately timbre-matched to the intended target speaker.

5. Conclusion

To address the challenges of achieving high-quality and robust speech conversion, we propose MPFM-VC, a novel model that utilizes ordinary differential equations (ODEs) to capture the dynamic evolution of speech features. A multi-dimensional perception mechanism is introduced to enhance the stability and naturalness of speech synthesis. In addition, a content perturbation-based training augmentation strategy is employed to improve the model's generalization ability by introducing controlled perturbations to content features during training, thereby reducing abnormal artifacts during inference. To further ensure accurate timbre matching, a voiceprint adversarial training strategy is adopted to explicitly disentangle content and speaker identity features, minimizing feature entanglement and residual speaker information. Experimental results demonstrate that the proposed algorithm, MPFM-VC, consistently outperforms existing state-of-the-art voice conversion models across multiple evaluation metrics, including MOS, MCD, and SECS, producing speech that is both more natural and speaker-consistent, with enhanced robustness and fidelity in real-world conditions.

6. Future work

In MPFM-VC, the current content perturbation-based training enhancement strategy primarily relies on the random masking of partial feature dimensions to improve the model's generalization capability. However, this approach may result in a slight degradation of speech quality and intelligibility. In future work, more advanced dynamic perturbation strategies—such as those based on adversarial generation or contrastive learning—could be explored to enhance the model's robustness against complex input noise and distributional shifts. Such approaches would enable the model to better adapt to the variability of real-world data, thereby improving its performance and stability in practical deployment scenarios.

Author Contributions: Y.W.; methodology, Y.W.; software, Y.W.; validation, Y.W.; formal analysis, Y.W.; investigation, Y.W.; resources, Y.W.; data curation, Y.W. and X.H.; writing—original draft preparation, Y.W., X.H ,S.L., T.Z., and Y.C.; writing—review and editing, Y.W.; visualization, X.H.; supervision, X.H.; project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research study received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are contained in this paper. Further inquiries (including but not limited to trained models and code) can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TTS	text to speech
VC	voice conversion
ODE	ordinary differential equation
MOS	Mean Opinion Score
MCD	Mel Cepstral Distortion
WER	Word Error Rate
SMOS	Similarity Mean Opinion Score
SECS	Speaker Embedding Cosine Similarity
ICASSP	IEEE International Conference on Acoustics, Speech and Signal Processing
ICLR	International Conference on Learning Representations
AAAI	Association for the Advancement of Artificial Intelligence Conference

References

1. Kaur, N.; Singh, P.. Conventional and contemporary approaches used in text to speech synthesis: A review. *Artificial Intelligence Review* **2023**, *56*, 5837–5880.
2. Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems* **2019**, *32*, .
3. Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv* 2020, preprint arXiv:2006.04558.
4. Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R.; J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; others, . Tacotron: Towards end-to-end speech synthesis. *arXiv* 2017, preprint arXiv:1703.10135.
5. Shen, J.; Pang, R.; Weiss, R.; J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; others, . Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, Location, 2018; 4779–4783.
6. Li, N.; Liu, S.; Liu, Y.; Zhao, S.; Liu, M.. Neural speech synthesis with transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, Location, 2019; 6706–6713.
7. Saito, Y.; Ijima, Y.; Nishida, K.; Takamichi, S.. Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Location, 2018; 5274–5278.
8. Qian, K.; Zhang, Y.; Chang, S.; Yang, X.; Hasegawa-Johnson, M.. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, Location, 2019; 5210–5219.
9. Chou, J.; Yeh, C.; Lee, H.. One-shot voice conversion by separating speaker and content representations with instance normalization. *arXiv* 2019, preprint arXiv:1904.05742.
10. Jia, Y.; Zhang, Y.; Weiss, R.; Wang, Q.; Shen, J.; Ren, F.; Nguyen, P.; Pang, R.; Lopez Moreno, I.; Wu, Y.; others, . Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems* **2018**, *31*, .

11. Radford, A.; Kim, J.; W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I.. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, Location, 2023; 28492–28518.
12. Qian, K.; Zhang, Y.; Gao, H.; Ni, J.; Lai, C.; Cox, D.; Hasegawa-Johnson, M.; Chang, S.. Contentvec: An improved self-supervised speech representation by disentangling speakers. In *International Conference on Machine Learning*, Location, 2022; 18003–18017.
13. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M.. Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* **2020**, *33*, 12449–12460.
14. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P.. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **1998**, *86*, 2278–2324.
15. Goodfellow, I.; J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y.. Generative adversarial nets. *Advances in neural information processing systems* **2014**, *27*, .
16. Rezende, D.; J.; Mohamed, S.; Wierstra, D.. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, Location, 2014; 1278–1286.
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; N.; Kaiser, L.; Polosukhin, I.. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*, .
18. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S.. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, Location, 2015; 2256–2265.
19. Lee, S.; Ping, W.; Ginsburg, B.; Catanzaro, B.; Yoon, S.. Bigvgan: A universal neural vocoder with large-scale training. *arXiv* 2022, preprint arXiv:2206.04658.
20. Kong, J.; Kim, J.; Bae, J.. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems* **2020**, *33*, 17022–17033.
21. Kaneko, T.; Tanaka, K.; Kameoka, H.; Seki, S.. ISTFTNet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time Fourier transform. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Location, 2022; 6207–6211.
22. Kim, J.; Kong, J.; Son, J.. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, Location, 2021; 5530–5540.
23. Kong, J.; Park, J.; Kim, B.; Kim, J.; Kong, D.; Kim, S.. Vits2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design. *arXiv* 2023, preprint arXiv:2307.16430.
24. Li, J.; Tu, W.; Xiao, L.. Freevc: Towards high-quality text-free one-shot voice conversion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Location, 2023; 1–5.
25. Lei, Y.; Yang, S.; Cong, J.; Xie, L.; Su, D.. Glow-wavegan 2: High-quality zero-shot text-to-speech synthesis and any-to-any voice conversion. *arXiv* 2022, preprint arXiv:2207.01832.
26. Pankov, V.; Pronina, V.; Kuzmin, A.; Borisov, M.; Usoltsev, N.; Zeng, X.; Golubkov, A.; Ermolenko, N.; Shirshova, A.; Matveeva, Y.. DINO-VITS: Data-Efficient Zero-Shot TTS with Self-Supervised Speaker Verification Loss for Noise Robustness. *arXiv* 2023, preprint arXiv:2311.09770.
27. Huang, W.; Violeta, L.; P.; Liu, S.; Shi, J.; Toda, T.. The singing voice conversion challenge 2023. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Location, 2023; 1–8.
28. Zhou, Y.; Chen, M.; Lei, Y.; Zhu, J.; Zhao, W.. VITS-based Singing Voice Conversion System with DSPGAN post-processing for SVCC2023. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Location, 2023; 1–8.
29. Ning, Z.; Jiang, Y.; Wang, Z.; Zhang, B.; Xie, L.. Vits-based singing voice conversion leveraging whisper and multi-scale f0 modeling. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Location, 2023; 1–8.
30. Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; Catanzaro, B.. Diffwave: A versatile diffusion model for audio synthesis. *arXiv* 2020, preprint arXiv:2009.09761.
31. Popov, V.; Vovk, I.; Gogoryan, V.; Sadekova, T.; Kudinov, M.. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International conference on machine learning*, Location, 2021; 8599–8608.
32. Vadim Popov, ; Ivan Vovk, ; Vladimir Gogoryan, ; Tasnima Sadekova, ; Mikhail Sergeevich Kudinov, ; Jiansheng Wei, . Diffusion-Based Voice Conversion with Fast Maximum Likelihood Sampling Scheme. In *International Conference on Learning Representations*, Location, 2022; .
33. Choi, H.; Lee, S.; Lee, S.. Dddm-vc: Decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Location, 2024; 17862–17870.

34. Ye, Z.; Xue, W.; Tan, X.; Chen, J.; Liu, Q.; Guo, Y.. Comospeech: One-step speech and singing voice synthesis via consistency model. In *Proceedings of the 31st ACM International Conference on Multimedia*, Location, 2023; 1831–1839.
35. Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; others, . Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, Location, 2024; .
36. Le, M.; Vyas, A.; Shi, B.; Karrer, B.; Sari, L.; Moritz, R.; Williamson, M.; Manohar, V.; Adi, Y.; Mahadeokar, J.; others, . Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems* **2023**, *36*, 14005–14034.
37. Guo, Y.; Du, C.; Ma, Z.; Chen, X.; Yu, K.. Voiceflow: Efficient text-to-speech with rectified flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Location, 2024; 11121–11125.
38. Guo, W.; Zhang, Y.; Pan, C.; Huang, R.; Tang, L.; Li, R.; Hong, Z.; Wang, Y.; Zhao, Z.. TechSinger: Technique Controllable Multilingual Singing Voice Synthesis via Flow Matching. *arXiv* 2025, preprint arXiv:2502.12572.
39. Du, J.; Lin, I.; Chiu, I.; Chen, X.; Wu, H.; Ren, W.; Tsao, Y.; Lee, H.; Jang, J.; R.. DFADD: The Diffusion and Flow-Matching Based Audio Deepfake Dataset. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, Location, 2024; 921–928.
40. Ronneberger, O.; Fischer, P.; Brox, T.. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, Location, 2015; 234–241.
41. Shi, Y.; Bu, H.; Xu, X.; Zhang, S.; Li, M.. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv* 2020, preprint arXiv:2010.11567.
42. Zhang, L.; Li, R.; Wang, S.; Deng, L.; Liu, J.; Ren, Y.; He, J.; Huang, R.; Zhu, J.; Chen, X.; others, . M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. *Advances in Neural Information Processing Systems* **2022**, *35*, 6914–6926.
43. An, K.; Chen, Q.; Deng, C.; Du, Z.; Gao, C.; Gao, Z.; Gu, Y.; He, T.; Hu, H.; Hu, K.; others, . Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv* 2024, preprint arXiv:2407.04051.
44. Wang, H.; Zheng, S.; Chen, Y.; Cheng, L.; Chen, Q.. Cam++: A fast and efficient network for speaker verification using context-aware masking. *arXiv* 2023, preprint arXiv:2303.00332.
45. Wei, H.; Cao, X.; Dan, T.; Chen, Y.. RMVPE: A robust model for vocal pitch estimation in polyphonic music. *arXiv* 2023, preprint arXiv:2306.15412.
46. Van Niekerk, B.; Carbonneau, M.; Zaïdi, J.; Baas, M.; Seuté, H.; Kamper, H.. A comparison of discrete and soft speech units for improved voice conversion. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Location, 2022; 6562–6566.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.