MF-Speech: Achieving Fine-Grained and Compositional Control in Speech Generation via Factor Disentanglement

Xinyue Yu[1], **Youqing Fang**[1], **Pingyu Wu**[1†], **Guoyang Ye**[1],
Wenbo Zhou[1], **Weiming Zhang**[1†], **Song Xiao**[2]

[1]School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China.
[2]Department of Electronic and Communication Engineering, Beijing Electronic Science and Technology Institute, Beijing 100070, China.
{xinyueyu, fangyq, wupingyu, studywithme}@mail.ustc.edu.cn,
{welbeckz, zhangwm}@ustc.edu.cn, xs_xidian@163.com

Abstract

Generating expressive and controllable human speech is one of the core goals of generative artificial intelligence, but its progress has long been constrained by two fundamental challenges: the deep entanglement of speech factors and the coarse granularity of existing control mechanisms. To overcome these challenges, we have proposed a novel framework called MF-Speech, which consists of two core components: MF-SpeechEncoder and MF-SpeechGenerator. MF-SpeechEncoder acts as a factor purifier, adopting a multi-objective optimization strategy to decompose the original speech signal into highly pure and independent representations of content, timbre, and emotion. Subsequently, MF-SpeechGenerator functions as a conductor, achieving precise, composable and fine-grained control over these factors through dynamic fusion and Hierarchical Style Adaptive Normalization (HSAN). Experiments demonstrate that in the highly challenging multi-factor compositional speech generation task, MF-Speech significantly outperforms current state-of-the-art methods, achieving a lower word error rate (WER=4.67%), superior style control (SECS=0.5685, Corr=0.68), and the highest subjective evaluation scores (nMOS=3.96, sMOS$_t$=3.86, sMOS$_e$=3.78). Furthermore, the learned discrete factors exhibit strong transferability, demonstrating their significant potential as a general-purpose speech representation.

Demo — https://guoyang25.github.io/mf-speech/
Extended version — https://arxiv.org/abs/2511.12074

Introduction

Infusing life into the digital world and endowing speech with personality and emotion is one of the most exciting frontiers in the field of generative artificial intelligence. From emotionally aware assistants to personalized voice restoration and expressive media synthesis (Sisman et al. 2020; Zhang et al. 2019; Veaux, Yamagishi, and King 2013), it is poised to transform how we interact with the digital world. Voice Conversion (Bargum, Serafin, and Erkut 2023) enables flexible manipulation of fundamental speech factors such as content, timbre, and emotion. As such, Voice Conversion has emerged as a key enabling technology toward this vision. However, two fundamental challenges have long troubled researchers in this field:

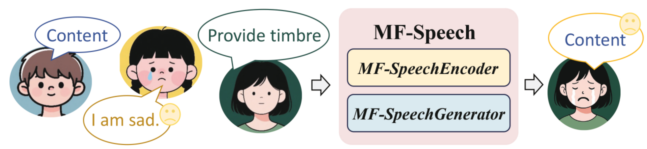

Figure 1: MF-Speech enables independent and fine-grained control over speech content, timbre, and emotion factors for speech synthesis.

- **Gene hybridization: The Challenge of Pure Factor Separation.** The content, timbre and emotion in speech are naturally intertwined and hard to separate. Due to the lack of a strong supervisory signal, the current strategies (Chou et al. 2018; Yadav et al. 2023; Qian et al. 2020b; Li, Han, and Mesgarani 2023; Wang et al. 2021b, 2018; Li, Han, and Mesgarani 2025; Chou, Yeh, and Lee 2019) act like rough filters, making it difficult to precisely and accurately separate various speech factors and thereby leading to timbre leakage and attribute interference. Moreover, this deep entanglement can also lead to fragile and chaotic factor representations, not only disrupting the precise control of each factor but also severely limiting their transferability across tasks (Li, Li, and Li 2023; Yuan et al. 2021; Lian, Zhang, and Yu 2022; Mu et al. 2024; Tu, Mak, and Chien 2024; Song et al. 2022; Deng et al. 2024).
- **Command failure: The Lack of Fine-Grained Control.** Even if pure speech factors are obtained, how to skillfully control them remains a major challenge. The existing control mechanisms are generally coarse-grained, just like using a sledgehammer to complete fine carving work. Whether models rely on the fundamental methods like static concatenation and implicit global modulation (Kaneko et al. 2019; Qian et al. 2019; Neekhara et al. 2023; Zhang, Ling, and Dai 2019; Yao et al. 2024), or employ advanced technologies such as dynamic fusion and explicit modulation (Yao et al. 2025; Ma et al. 2024; Ning et al. 2023; Choi and Park 2025; Li

[†]Corresponding author.

et al. 2024a; Qian et al. 2020a), they consistently suffer from coarse-grained control. This is due to their failure to systematically combine dynamic weights and hierarchical style injection. Therefore, models often struggle to balance content fidelity (content) and style similarity (timbre and emotion). This fundamental defect cannot be remedied through post-processing techniques (Ren et al. 2020; Lin et al. 2024; Tian, Liu, and Lee 2024).

To address these two major challenges, we propose MF-Speech, a framework designed to achieve fine-grained and compositional control in speech generation via multi-factor disentanglement (Figure 1). This framework consists of **Multi-factor Speech Encoder (MF-SpeechEncoder)** and **Multi-factor Speech Generator (MF-SpeechGenerator)**. It fundamentally addresses the aforementioned challenges by enhancing the purification capability and clarifying the command direction, achieving composable and fine-grained control speech generation. Our main contributions can be summarized as follows:

- **Multi-factor Speech Encoder to ensure factor purity (MF-SpeechEncoder).** We designed a speech factor purifier that uses a three-stream architecture and decomposes the raw speech signal into three highly pure and mutually independent information streams: content, timbre, and emotion. This ensures a high degree of independence for subsequent control and addresses the challenge of gene hybridization.
- **Multi-factor Speech Generator to enhance control granularity (MF-SpeechGenerator).** Building upon the purified factors, we developed the speech factor conductor. This component moves beyond coarse control by incorporating dynamic fusion and Hierarchical Style Adaptive Normalization (HSAN). This enables highly fine-grained control over timbre and emotion. As a result, the model can synthesize a vast array of speech combinations with high style similarity, while maintaining content fidelity.
- **Comprehensive empirical and systematic validation.** Extensive experiments demonstrate the effectiveness of our proposed framework. Results show that MF-SpeechEncoder can effectively purify speech factors to ensure control independence. Moreover, in the challenging task of multi-factor compositional speech generation, the MF-Speech demonstrates remarkable controllability in terms of content fidelity and style similarity.

Related Work

Factor Disentanglement: Strategies and Challenges. VQMIVC (Wang et al. 2021a) separates pitch, content, and timbre through vector quantization and mutual information minimization. However, F0 is not explicitly modeled but rather provided as an external condition, leading to unnatural prosody in converted speech. Moreover, as pitch is a fundamental acoustic feature, it is inherently entangled with timbre, resulting in implicit pitch–timbre entanglement and residual timbre leakage. StyleVC (Hwang, Lee, and Lee 2022) uses a global style encoder that outputs a single vector, making it difficult to disentangle timbre and emotion. Prosodic information is only supplemented through external auxiliary features, and adversarial training is limited to content–style separation, allowing emotion and timbre to remain entangled within the style representation. StableVC (Yao et al. 2025) incorporates a gradient reversal layer (GRL) on top of FaCodec's style component to further disentangle style and timbre factors. However, since FaCodec is primarily optimized for F0 modeling, emotion factor remains implicitly entangled and is not explicitly modeled. In summary, existing methods often suffer from impure factor definitions, limited architectural design, and limited training objectives, resulting in incomplete disentanglement. In contrast, we explicitly model emotion factor based on prosody information and design dedicated modules for timbre, emotion, and content under a unified multi-objective optimization framework. This enables reduced mutual interference among the factors and enhances the ability to control them independently.

Conditional Generation: Mechanisms and Granularity. MSM-VC (Wang et al. 2023) and DDDM-VC (Choi, Lee, and Lee 2024) employ static feature concatenation and inject it into the decoder, which offers limited flexibility in style control. NS2VC (Shen et al. 2023) uses a FiLM-based affine transformation driven by speech prompts, achieving a degree of dynamic control, yet its mechanism remains less flexible than a fully disentangled, hierarchical design. Facodec (Ju et al. 2024) simulates dynamic control via serial prompt concatenation, but this essentially functions as an advanced form of static fusion, lacking real-time adjustment of factor weights. Although StableVC (Yao et al. 2025) leverages conditional flow matching and attention mechanisms to improve dynamic feature fusion—thereby enhancing the independent control of timbre and style—it still relies on relatively static generation control with a shallow overall structure and lacks explicit dynamic weighting of factor contributions. HierVST (Lee et al. 2023) presents a hierarchical style injection mechanism, a structural advancement, but it lacks dynamic factor weighting and explicit balancing of factor contributions. In summary, existing control mechanisms are typically designed in a static, global, or non-collaborative manner, which limits their capacity for fine-grained and compositional control. In contrast, our framework employs a dynamic fusion module and HSAN. This approach not only considers the adaptive weighting of each factor but also enables style injection in each layer in a fine-grained manner, significantly enhancing the flexibility of multi-factor composition.

MF-Speech

Overview

To enable more independent, fine-grained, and controllable speech synthesis, we propose MF-Speech (Figure 2). It extracts three independent and high-purity speech factors. Then, it dynamically integrates these factors and injects the timbre and emotion at each layer. Thereby, it improves the content fidelity and style similarity of the generated speech. MF-Speech consists of two core components:

- **MF-SpeechEncoder**: A high-purity factor encoder,

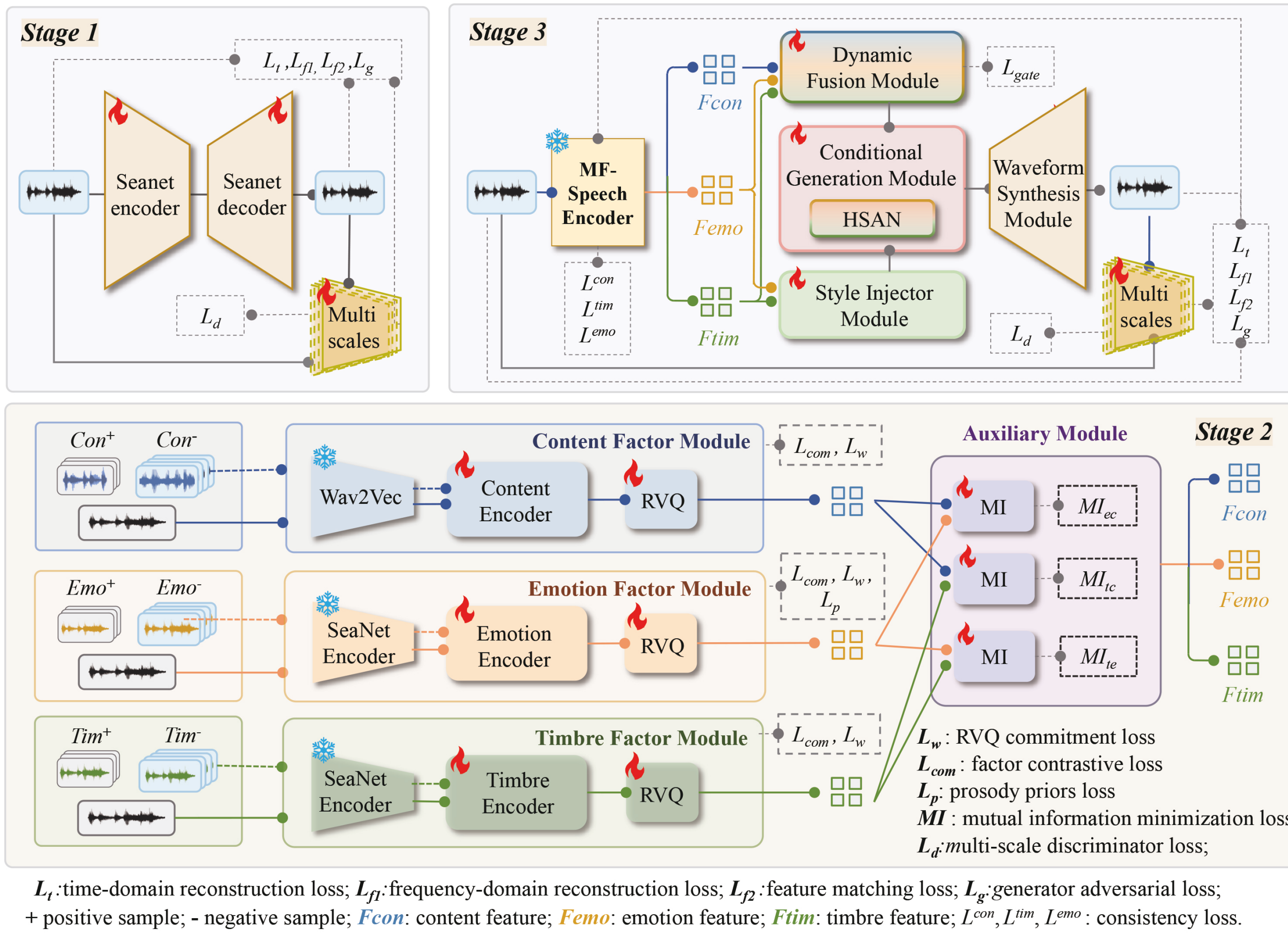

Figure 2: The training process of MF-Speech consists of three stages. The first stage ensures high-precision conversion between waveforms and features. The second stage disentangles clean and independent content, timbre, and emotion factors. The third stage enables fine-grained, multi-factor control for waveform generation.

which extracts disentangled discrete representations for each core factor from input speech.

- **MF-SpeechGenerator**: A fine-grained waveform generator, which synthesizes the final speech by precisely controlling the fusion and modulation based on arbitrary combinations of discrete factor representations.

MF-SpeechEncoder

The design of MF-SpeechEncoder aims to learn high-purity and mutually independent representations of speech content, timbre, and emotion, which is an essential foundation for fine-grained and compositional control in speech generation. As illustrated in Figure 2, MF-SpeechEncoder is trained in stage 2 under a multi-objective optimization strategy. The framework adopts a three-stream architecture, consisting of three specialized submodules and an information-theoretic auxiliary module.

- **Content Factor Module:** To isolate pure speech content, this module first extracts initial representations using a pre-trained Wav2Vec2 (Baevski et al. 2020) model as its backbone. A lightweight trainable sub-network then refines these representations via sentence-level content contrastive learning, specifically to suppress residual timbre and emotion information. Finally, a Residual Vector Quantizer (RVQ) (Yang et al. 2023) discretizes these purified representations into discrete content tokens.
- **Emotion Factor Module:** Recognizing that emotional expression heavily relies on prosody dynamics, this module adopts a two-stage architecture. Initially, lightweight predictors explicitly generate F0 and energy representations from an intermediate layer, guided by direct supervision to focus on emotion-related acoustic cues. The final emotion representation is then derived from these predicted prosody representations. An emotion contrastive loss enhances the discriminability of these representations across different emotional states, which are subsequently discretized using RVQ to yield controllable and transferable emotion representations.
- **Timbre Factor Module:** This module aims to create stable and generalizable timbre representations. After a

SeaNet encoder (Tagliasacchi et al. 2020), it employs a multi-head attention mechanism (Deora et al. 2023) to aggregate and enhance global timbre representations from the input speech. To further purify these representations from content and emotion interference and bolster representation robustness, a timbre-specific contrastive loss is applied. The resulting representations are then discretized using RVQ.

- **Information Theory Constraints:** To mitigate potential residual entanglement between the outputs of the dedicated factor modules, we apply structural regularization using mutual information (MI) minimization after discretization. Inspired by MAIN-VC (Li et al. 2024b), a separate MI estimation network employing CLUB (Cheng et al. 2020) and MINE (Belghazi et al. 2018) are trained to penalize redundant information between the factor representations, thereby promoting their independence. This MI estimation network is active only during training to maintain operational efficiency at inference.
- **Optimization Objectives:** The MF-SpeechEncoder is trained to produce disentangled discrete representations of content, timbre, and emotion using a multi-objective loss function. This total loss, $\mathcal{L}_{\text{Encoder}}$, combines objectives for RVQ $\mathcal{L}_w^f$, factor-specific contrastive learning $\mathcal{L}_{\text{com}}$, prosody priors $\mathcal{L}_{\text{p}}$, and mutual information (MI) minimization constraints $\mathcal{L}_{\text{MI}}$. The MI constraints are introduced gradually via a warm-up schedule to avoid hindering initial representation learning. The overall MF-SpeechEncoder objective is shown in Equation 1, where t, e, c represent content, timbre, and emotion factors, respectively, and α(epoch) is the warm-up weighting for MI loss. See the *Appendix A* for more details.

$$\begin{aligned}\mathcal{L}_{\text{Encoder}} = &\sum_{f\in\{\text{t,e,c}\}} \lambda_{com}^f \cdot \mathcal{L}_{\text{com}}^f + \sum_{f\in\{\text{t,e,c}\}} \lambda_w^f \cdot \mathcal{L}_w^f \\ &+ \lambda_p \cdot \mathcal{L}_{\text{p}} + \alpha(\text{epoch}) \cdot \sum_{\text{X,Y}} \mathcal{L}_{\text{MI}}(\text{X,Y}).\end{aligned} \tag{1}$$

MF-SpeechGenerator

MF-SpeechGenerator performs fine-grained and compositional control in Speech Generation based on the discrete factor representations provided by MF-SpeechEncoder, through dynamic fusion and HSAN. As shown in Figure 2, this process is trained in stage 3, which comprises four collaboratively functioning modules: dynamic fusion, style injection, conditional generation, and waveform synthesis. Dynamic fusion and HSAN are shown in Figure 3.

- **Dynamic Fusion Module:** This initial module dynamically integrates the discrete content, timbre, and emotion representations from the MF-SpeechEncoder into a unified conditional representation for subsequent acoustic modeling. It employs a dynamic gating mechanism that generates time-varying weights for each factor. This adaptive, weighted integration allows the model to flexibly regulate the influence of each factor at different time steps during synthesis.
- **Style Injection Module:** To enable fine-grained control over timbre and emotion, this module is essentially a style parameter generator. It infers multi-level style parameter representations from the discrete representations of timbre and emotion. This facilitates their integration into the generator to implement the HSAN mechanism.

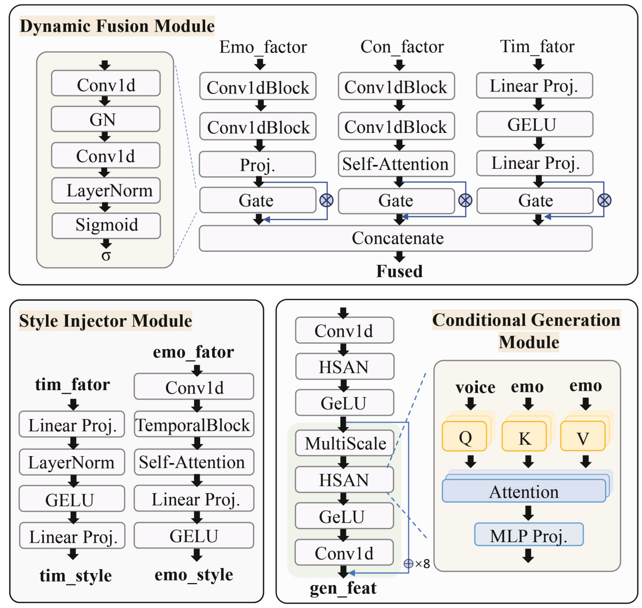

Figure 3: The related architectures and data flows of dynamic fusion and HSAN

- **Conditional Generation Module:** This module constructs highly conditional temporal acoustic representations. It maps the dynamically fused factor representations into detailed acoustic representation sequences, guided by the hierarchical style parameters from the style parameter generator. The backbone network uses stacked residual blocks integrated with multi-scale convolution modules to effectively capture complex temporal dependencies. Crucially, the HSAN layer is applied across multiple levels of this network. At each level, style parameters adaptively modulate the representations, ensuring the acoustic generation process is consistently conditioned and precisely controlled by fine-grained style information. HSAN first fuses timbre and emotion representations via cross-attention. This fused result is projected to yield affine parameters (γ, β) and a residual modulation term (α). Equation 2 defines a transformation that combines an affine function with residual modulation, where $IN(x)$ denotes instance normalization, x is the input feature, and λ is a scalar. This formulation allows normalized features to be scaled and shifted, with an added expressive residual term enhancing stylistic control.

$$\mathbf{y} = \text{IN}(\mathbf{x})(1 + \tanh(\boldsymbol{\gamma})) + \boldsymbol{\beta} + \lambda \tanh(\alpha) \odot \mathbf{x}. \tag{2}$$

- **Waveform Synthesis Module:** Finally, this module converts the generated acoustic representation sequence into the speech waveform. We employ the SeaNet decoder (Tagliasacchi et al. 2020) as the waveform decoder. Its robust capability ensures preservation of acoustic details and high perceptual quality. During training, it is initially frozen and then fine-tuned via a phased unfreezing strategy, allowing it to adapt optimally to the generator's output and further enhance synthesis quality.

- **Optimization Objectives:** The MF-SpeechGenerator is trained using adversarial learning, with a phased unfreezing strategy for the pre-trained waveform decoder, to enhance realism and ensure stable convergence. The generator's composite loss function ($\mathcal{L}_{\text{Generator}}$) is shown in Equation 3. See the *Appendix B* for more details.

$$\begin{aligned}\mathcal{L}_{\text{Generator}} = \lambda_{\text{gate}}\mathcal{L}_{\text{gate}} + \lambda_g\mathcal{L}_g + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}} \\ + \lambda_t\mathcal{L}_t + \lambda_f\mathcal{L}_f + \lambda_{\text{sim}}\mathcal{L}_{\text{sim}}.\end{aligned} \tag{3}$$

The multi-scale discriminator (Défossez et al. 2022) loss ($\mathcal{L}_d$) is shown in Equation 4. It is a hinge loss to distinguish real speech x from generated speech $\hat{x}$.

$$\mathcal{L}_d = \frac{1}{K}\sum_{k=1}^{K}\left[\max(0, 1 - D_k(x)) + \max(0, 1 + D_k(\hat{x}))\right]. \tag{4}$$

Experiments

Experimental Setup

Dataset and Training Details: Since the ESD dataset (Zhou et al. 2021) contains explicit emotion labels, we selected it for our experiments and divided it into a training set, a seen test set, and an unseen test set. The MF-Speech framework was trained in three stages on a single NVIDIA 4090 GPU. Stage 1 was trained for 92,000 iterations with a batch size of 24, λ_g=3, λ_{feat}=3, λ_t=0.1, λ_f=1. Stage 2 was trained for 27,500 iterations with a batch size of 12, λ_{com}=5, λ_w=1, λ_p=2. Stage 3 was trained for 91,800 iterations with a batch size of 72, λ_{gate}=1, λ_{sim}=1, λ_g=3 , λ_{feat}=3, λ_t=0.1, λ_f=1.

Baseline Systems: To benchmark the effectiveness of MF-Speech in fine-grained and compositional control in speech generation, we select four representative baselines that cover the main paradigms in this field. **StyleVC** (Hwang, Lee, and Lee 2022) leverages adversarial training to achieve disentanglement between content and style, demonstrating strong adaptability to non-parallel data. **NS2VC** (Shen et al. 2023) introduces a latent diffusion mechanism to implicitly model global style representations. **FaCodec** (Ju et al. 2024) employs explicit factor decomposition, offering high structural interpretability and fine-grained control at the factor level. **DDDM-VC** (Choi, Lee, and Lee 2024) adopts an iterative diffusion process with multi-modal conditioning to enable fine-grained and flexible style control. By comparing with these baselines, we can verify MF-Speech's independence and fine-grained control capabilities.

Evaluation Overview

Evaluation Scheme: The input data is constructed as content (c), timbre (t), and emotion (e). When $c = e = t$, it corresponds to the speech reconstruction task, generating reconstruction data. If any one of them differs, it corresponds to the multi-factor compositional speech generation task, producing controllable data. Using this setup, we generated 200 reconstruction samples and 200 controllable samples with both the MF-Speech and baseline methods, and evaluated the task performance accordingly. To evaluate the independence of each factor, we assess performance on target and non-target tasks, compute mutual information between factor pairs, and use t-SNE to visualize the factor representations. The specific calculation details of the indicators are in *Appendix C*.

Subjective Evaluation: We selected 20 participants for subjective evaluation. **nMOS:** Overall perceived naturalness of the generated speech. **sMOS$_t$, sMOS$_e$:** Similarity of timbre (sMOS$_t$) and emotion (sMOS$_e$) to reference speech.

Objective Evaluation. MI: Quantify redundancy between the learned discrete representations of content (c), timbre (t), and emotion (e). **Acc:** Accuracy on target tasks (e.g., speaker ID from timbre) and non-target tasks (e.g., emotion ID from timbre) to measure information leakage. **t-SNE:** Used to qualitatively illustrate the clustering and separability of the learned factor representations. **UTMOS (Saeki et al. 2022):** An objective MOS prediction system. **SECS:** Cosine similarity between speaker embeddings of generated and reference samples. **Log RMSE, Corr:** Log Root Mean Squared Error (Log RMSE) and the Pearson Correlation Coefficient (Corr) for F0 contours, aligned using Dynamic Time Warping (DTW) (Müller 2007). These are used to evaluate style similarity, as emotion is highly correlated with F0. **WER:** Word Error Rate, calculated by comparing ground truth transcriptions with those from a pre-trained Automatic Speech Recognition (ASR) model applied to the generated speech.

Experimental Results on MF-Speech

We evaluated MF-Speech on the speech reconstruction task and the multi-factor compositional speech generation task, with detailed results presented in Table 1. Overall, MF-Speech demonstrates strong control, as measured by metrics for content fidelity and style similarity.

Speech Reconstruction: MF-Speech achieves the highest timbre similarity (SECS = 0.7401) and the lowest word error rate (WER = 2.83%), demonstrating strong control over both timbre and content. Its emotion consistency is also competitive, with a high F0 correlation (Corr = 0.94) and a low F0 reconstruction error (Log RMSE = 0.08), closely approaching the performance of the best-performing model, DDDM-VC. In terms of subjective naturalness (nMOS = 3.53) and style similarity (sMOS$_t$ = 3.54; sMOS$_e$ = 3.50), MF-Speech performs comparably to StyleVC and is only slightly behind DDDM-VC. The perceptual quality of the generated speech may have influenced similarity judgments to some extent. Overall, while DDDM-VC slightly outperforms MF-Speech across several metrics, the differences are relatively minor. Notably, the factor controllability and disentanglement demonstrated in the reconstruction task remain limited.

Multi-factor Compositional Speech Generation: To further evaluate controllability, we conducted a more challenging ex periment on multi-factor compositional speech generation, where content, timbre, and emotion are combined from different sources. Apart from UTMOS, MF-

Speech Reconstruction								
Name	nMOS↑	sMOS$_t$↑	sMOS$_e$↑	UTMOS↑	SECS ↑	Log RMSE↓	Corr↑	WER↓
GT	-	-	-	3.9193	-	-	-	-
StyleVC	3.54 ± 0.15	3.56 ± 0.14	3.32 ± 0.16	3.4536	0.2015	0.23	0.5	28.16%
NS2VC	2.67 ± 0.17	2.52 ± 0.18	2.63 ± 0.17	2.3982	0.2758	0.12	0.87	7.17%
DDDM-VC	**3.87 ± 0.11**	**3.86 ± 0.12**	**3.79 ± 0.10**	**3.6116**	0.6430	**0.03**	**0.99**	5.17%
FACodec	2.30 ± 0.15	2.16 ± 0.16	2.27 ± 0.15	1.9146	0.3002	0.09	0.92	18.17%
MF-Speech	3.53 ± 0.13	3.54 ± 0.13	3.50 ± 0.13	2.7314	**0.7401**	0.08	0.94	**2.83%**
Multi-factor Compositional Speech Generation								
Name	nMOS↑	sMOS$_t$↑	sMOS$_e$↑	UTMOS↑	SECS ↑	Log RMSE↓	Corr↑	WER↓
StyleVC	2.81 ± 0.31	2.98 ± 0.33	2.40 ± 0.36	**3.2176**	0.0985	0.35	0.48	24.83%
NS2VC	3.76 ± 0.26	3.11 ± 0.31	3.44 ± 0.25	2.0270	0.1552	0.43	0.55	23.33%
DDDM-VC	3.58 ± 0.34	3.50 ± 0.26	3.13 ± 0.33	2.8388	0.3723	0.37	0.62	11.67%
FACodec	2.83 ± 0.34	2.38 ± 0.34	3.14 ± 0.30	1.7128	0.1866	0.41	0.58	29.17%
MF-Speech	**3.96 ± 0.31**	**3.86 ± 0.30**	**3.78 ± 0.27**	2.6686	**0.5685**	**0.34**	**0.68**	**4.67%**

Table 1: The subjective and objective evaluation results of MF-Speech and baseline systems and "GT" refers to the real sample.

Speech achieved the best performance across all other metrics (SECS = 0.5685, Log RMSE = 0.34, Corr = 0.68, and WER = 4.67%). Although MF-Speech achieves a slightly lower UTMOS score than StyleVC, the noticeably faster speaking rate of StyleVC negatively affects its perceived naturalness, resulting in a higher subjective evaluation score for MF-Speech. These results demonstrate that MF-Speech can successfully generate speech with accurate content and high similarity in both timbre and emotion, highlighting its strong capability in controllable and truly compositional generation.

Summary of MF-Speech Performance: These results clearly demonstrate that MF-Speech achieves fine-grained and compositional speech generation. The generated speech has high fidelity (content) and strong style similarity (timbre and emotion). Although slightly behind DDDM-VC in reconstruction, MF-Speech excels in multi-factor compositional generation, showing strong control over content, timbre, and emotion.

Experimental Results on MF-SpeechEncoder

In order to further assess the high-purity and independence of the factors, we evaluated the results of MF-SpeechEncoder, as detailed in Table 2. In terms of mutual information, the performance is comparable to that of the baselines. However, in terms of the results of the target tasks (0.9979%, 0.9296%, 0.9593%) and non-target tasks (0.2618%, 0.0054%, 0.1421%), MF-SpeechEncoder has a more significant advantage. Although Facodec shows advantages in some metrics, MF-SpeechEncoder's t-SNE visualizations (Figure 4.(a)-(d)) clearly demonstrate superior cluster separability and compactness, confirming robust disentanglement. Therefore, from multiple perspectives, MF-SpeechEncoder can produce more independent and pure factor representations. This effective separation of factors is foundational to MF-Speech's fine-grained and compositional control.

Metrics	StyleVC	DDDM-VC	Facodec	**MF-SE**
MI$_{te}$↓	0.0070	0.0080	**0.0063**	0.0076
MI$_{tc}$↓	0.0063	0.0079	**0.0059**	0.0061
MI$_{ec}$↓	0.0080	**0.0052**	0.0197	0.0061
Acc$_t$↑	0.9861	0.9882	0.9939	**0.9979**
Acc$_e$↑	0.5168	0.2075	0.2343	**0.9296**
Acc$_c$↑	0.6661	0.8789	0.0068	**0.9593**
Acc$_{te}$↓	0.5400	0.5579	0.5932	**0.2618**
Acc$_{tc}$↓	0.0321	0.0207	0.0907	**0.0054**
Acc$_{et}$↓	0.7900	**0.1339**	0.1596	0.2486
Acc$_{ec}$↓	0.1111	0.0054	**0.0021**	0.0089
Acc$_{ct}$↓	0.4268	0.9871	0.1704	**0.1421**
Acc$_{ce}$↓	0.4861	0.8118	**0.2332**	0.2529

Table 2: The objective evaluation results of disentanglement. "MF-SE" denotes the MF-SpeechEncoder.

Ablation Study

We conducted ablation studies on key components in the MF-SpeechEncoder and MF-SpeechGenerator modules separately to validate their respective contributions.

MF-SpeechEncoder Component Analysis: We ablated three components from the full MF-SpeechEncoder: Mutual Information constraints (M1: w/o_{MI}), contrastive learning (M2: w/o_{com}), and prosody priors (M3: w/o_{pro}). **MI Constraints (w/o_{MI} - M1)(Figure 4.(f), (j)):** After removing the MI constraint, the clustering boundaries of the emotion t-SNE become blurred, and the compactness of the clustering of timbre and emotion also decreases. This highlights the crucial role of MI in enhancing the discrimination and structural integrity of the factors. **Contrastive Learning (w/o_{com} - M2)(Figure 4.(g), (k)):** The removal of contrastive learning directly led to severe information entanglement. The t-SNE plots of timbre and emotion showed significant overlap and lacked clear boundaries when contrastive learning was removed, confirming that it plays a crucial role in achieving effective disentanglement. This observation suggests that minimizing mutual information between branches alone is insufficient; it is also essential to impose

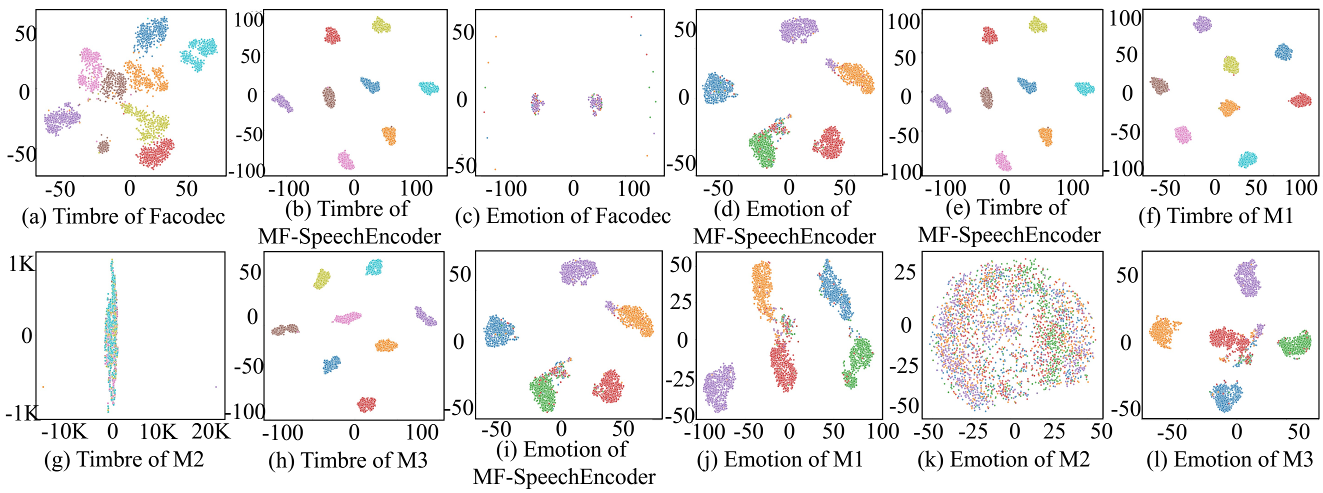

Figure 4: t-SNE visualization of timbre and emotion representation from four models.

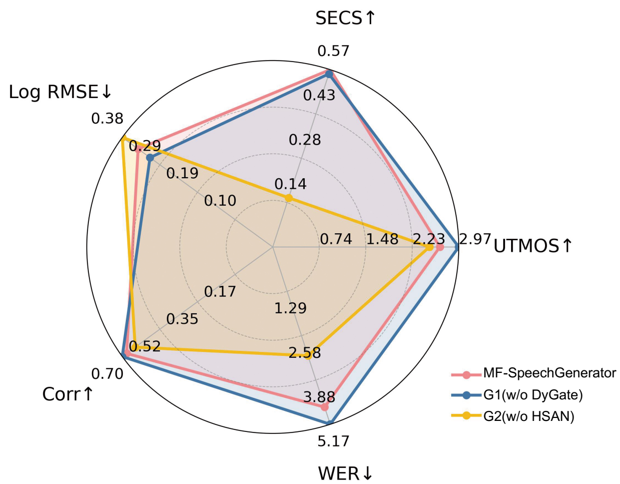

Figure 5: The ablation results of SpeechGenerator.

explicit constraints on each individual branch to preserve their respective factor-specific information. **Prosody Priors** (w/o_{pro} **- M3)(Figure 4.(h), (l)):** Removing prosody priors impaired timbre clustering and led to disordered emotion clusters. This highlights the crucial importance of prosody priors for robust emotion modeling. As emotion representations become more accurately extracted, the mutual information constraints enable better timbre disentanglement. Therefore, MF-SpeechEncoder demonstrates high purity and stability in disentanglement, and its performance consistently outperforms all ablated variants.

SpeechGenerator Component Analysis: We evaluated the impact of the dynamic fusion module (G1: (w/o_{DyGate})) and the HSAN (G2: (w/o_{HSAN})) on multi-factor compositional speech generation, with results in Figure 5. **Dynamic Fusion** ((w/o_{DyGate}) **- G1):** Removing dynamic fusion slightly reduced timbre similarity (SECS from 0.5685 to 0.5551) and increased the word error rate (WER from 4.76% to 5.17%). Although it showed marginal improvements in UTMOS, Log RMSE, and Corr, the differences were minimal. Given that WER and SECS are more critical for evaluating controllability, these results suggest that dynamic fusion plays a positive role in preserving the consistency of both timbre and content. **HSAN** ((w/o_{HSAN}) **- G2):** Removing HSAN had a more pronounced negative impact. Except for WER, all other metrics deteriorated significantly (SECS dropped to 0.1576, Corr dropped to 0.64, and Log RMSE increased to 0.38). The relatively better WER can be attributed to the fact that, without HSAN, the fused features are used directly for generation, eliminating the need for additional control injections of timbre and emotion. This allows the model to focus more on content representation, thereby improving WER. However, our primary focus is on fine-grained and compositional control. Despite the improved WER, the ability to control timbre and emotion becomes significantly weaker without HSAN. Therefore, HSAN remains an indispensable component, playing a crucial role in ensuring style consistency and controllable synthesis. Overall, dynamic fusion and HSAN play complementary and crucial roles: dynamic fusion enables multi-factor coordination and content integrity, while HSAN is key to powerful style expression and control.

Conclusion

We present MF-Speech, a framework that tackles factor entanglement and coarse controllability in speech generation. It contains two main components: MF-SpeechEncoder, a factor purifier that uses a multi-objective scheme to produce highly disentangled discrete representations of content, timbre, and emotion; and MF-SpeechGenerator, a speech conductor that fuses these purified factors via dynamic fusion and HSAN for fine-grained, compositional control. Experiments show that MF-Speech outperforms prior methods in content fidelity and overall style similarity. In future, we will further enhance synthesis quality and generalization.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62121002,62372423,62072421,62476013 and was also supported by the Fundamental Research Funds for the Central Universities WK2100250070

References

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 12449–12460.

Bargum, A. R.; Serafin, S.; and Erkut, C. 2023. Reimagining speech: A scoping review of deep learning-powered voice conversion. *arXiv preprint arXiv:2311.08104*.

Belghazi, M. I.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, R. D. 2018. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*.

Cheng, P.; Hao, W.; Dai, S.; Liu, J.; Gan, Z.; and Carin, L. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, 1779–1788. PMLR.

Choi, H.-Y.; Lee, S.-H.; and Lee, S.-W. 2024. 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*, volume 38, 17862–17870.

Choi, H.-Y.; and Park, J. 2025. VoicePrompter: Robust Zero-Shot Voice Conversion with Voice Prompt and Conditional Flow Matching. *arXiv preprint arXiv:2501.17612*.

Chou, J.-c.; Yeh, C.-c.; and Lee, H.-y. 2019. One-shot voice conversion by separating speaker and content representations with instance normalization. *arXiv preprint arXiv:1904.05742*.

Chou, J.-c.; Yeh, C.-c.; Lee, H.-y.; and Lee, L.-s. 2018. Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations. *arXiv preprint arXiv:1804.02812*.

Défossez, A.; Copet, J.; Synnaeve, G.; and Adi, Y. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Deng, Y.; Tang, H.; Zhang, X.; Cheng, N.; Xiao, J.; and Wang, J. 2024. Learning disentangled speech representations with contrastive learning and time-invariant retrieval. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7150–7154. IEEE.

Deora, P.; Ghaderi, R.; Taheri, H.; and Thrampoulidis, C. 2023. On the optimization and generalization of multi-head attention. *arXiv preprint arXiv:2310.12680*.

Hwang, I.-S.; Lee, S.-H.; and Lee, S.-W. 2022. StyleVC: non-parallel voice conversion with adversarial style generalization. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 23–30. IEEE.

Ju, Z.; Wang, Y.; Shen, K.; Tan, X.; Xin, D.; Yang, D.; Liu, Y.; Leng, Y.; Song, K.; Tang, S.; et al. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*.

Kaneko, T.; Kameoka, H.; Tanaka, K.; and Hojo, N. 2019. Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion. *arXiv preprint arXiv:1907.12279*.

Lee, S.-H.; Choi, H.-Y.; Oh, H.-S.; and Lee, S.-W. 2023. Hiervst: Hierarchical adaptive zero-shot voice style transfer. *arXiv preprint arXiv:2307.16171*.

Li, D.; Li, X.; and Li, X. 2023. DVQVC: An unsupervised zero-shot voice conversion framework. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Li, J.; Guo, Y.; Chen, X.; and Yu, K. 2024a. Sef-vc: Speaker embedding free zero-shot voice conversion with cross attention. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12296–12300. IEEE.

Li, P.; Wang, J.; Zhang, X.; Zhang, Y.; Xiao, J.; and Cheng, N. 2024b. MAIN-VC: Lightweight Speech Representation Disentanglement for One-shot Voice Conversion. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–7. IEEE.

Li, Y. A.; Han, C.; and Mesgarani, N. 2023. Styletts-vc: One-shot voice conversion by knowledge transfer from style-based tts models. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, 920–927. IEEE.

Li, Y. A.; Han, C.; and Mesgarani, N. 2025. Styletts: A style-based generative model for natural and diverse text-to-speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*.

Lian, J.; Zhang, C.; and Yu, D. 2022. Robust disentangled variational speech representation learning for zero-shot voice conversion. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6572–6576. IEEE.

Lin, W.; He, C.; Mak, M.-W.; Lian, J.; and Lee, K. A. 2024. VoxGenesis: Unsupervised Discovery of Latent Speaker Manifold for Speech Synthesis. *arXiv preprint arXiv:2403.00529*.

Ma, L.; Zhu, X.; Lv, Y.; Wang, Z.; Wang, Z.; He, W.; Zhou, H.; and Xie, L. 2024. Vec-Tok-VC+: Residual-enhanced Robust Zero-shot Voice Conversion with Progressive Constraints in a Dual-mode Training Strategy. *arXiv preprint arXiv:2406.09844*.

Mu, Z.; Yang, X.; Sun, S.; and Yang, Q. 2024. Self-supervised disentangled representation learning for robust target speech extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18815–18823.

Müller, M. 2007. Dynamic Time Warping. In *Information Retrieval for Music and Motion*, 69–84. Springer.

Neekhara, P.; Hussain, S.; Valle, R.; Ginsburg, B.; Ranjan, R.; Dubnov, S.; Koushanfar, F.; and McAuley, J. 2023. SelfVC: Voice conversion with iterative refinement using self transformations. *arXiv preprint arXiv:2310.09653*.

Ning, Z.; Xie, Q.; Zhu, P.; Wang, Z.; Xue, L.; Yao, J.; Xie, L.; and Bi, M. 2023. Expressive-vc: Highly expressive voice conversion with attention fusion of bottleneck and perturbation features. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Qian, K.; Jin, Z.; Hasegawa-Johnson, M.; and Mysore, G. J. 2020a. F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6284–6288. IEEE.

Qian, K.; Zhang, Y.; Chang, S.; Hasegawa-Johnson, M.; and Cox, D. 2020b. Unsupervised speech decomposition via triple information bottleneck. In *International Conference on Machine Learning*, 7836–7846. PMLR.

Qian, K.; Zhang, Y.; Chang, S.; Yang, X.; and Hasegawa-Johnson, M. 2019. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, 5210–5219. PMLR.

Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Saeki, T.; Xin, D.; Nakata, W.; Koriyama, T.; Takamichi, S.; and Saruwatari, H. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Shen, K.; Ju, Z.; Tan, X.; Liu, Y.; Leng, Y.; He, L.; Qin, T.; Zhao, S.; and Bian, J. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*.

Sisman, B.; Yamagishi, J.; King, S.; and Li, H. 2020. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 132–157.

Song, W.; Yue, Y.; Zhang, Y.-j.; Zhang, Z.; Wu, Y.; and He, X. 2022. Multi-Speaker Multi-Style speech synthesis with timbre and style disentanglement. In *National Conference on Man-Machine Speech Communication*, 132–140. Springer.

Tagliasacchi, M.; Li, Y.; Misiunas, K.; and Roblek, D. 2020. SEANet: A multi-modal speech enhancement network. *arXiv preprint arXiv:2009.02095*.

Tian, Y.; Liu, J.; and Lee, T. 2024. User-Driven Voice Generation and Editing through Latent Space Navigation. *arXiv preprint arXiv:2408.17068*.

Tu, Y.; Mak, M.-W.; and Chien, J.-T. 2024. Contrastive self-supervised speaker embedding with sequential disentanglement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Veaux, C.; Yamagishi, J.; and King, S. 2013. Towards personalised synthesised voices for individuals with vocal disabilities: Voice banking and reconstruction. In *Proceedings of the fourth workshop on speech and language processing for assistive technologies*, 107–111.

Wang, D.; Deng, L.; Yeung, Y. T.; Chen, X.; Liu, X.; and Meng, H. 2021a. Vqmivc: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion. *arXiv preprint arXiv:2106.10132*.

Wang, J.; Li, J.; Zhao, X.; Wu, Z.; Kang, S.; and Meng, H. 2021b. Adversarially learning disentangled speech representations for robust multi-factor voice conversion. *arXiv preprint arXiv:2102.00184*.

Wang, Y.; Stanton, D.; Zhang, Y.; Ryan, R.-S.; Battenberg, E.; Shor, J.; Xiao, Y.; Jia, Y.; Ren, F.; and Saurous, R. A. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International conference on machine learning*, 5180–5189. PMLR.

Wang, Z.; Wang, X.; Xie, Q.; Li, T.; Xie, L.; Tian, Q.; and Wang, Y. 2023. Msm-vc: High-fidelity source style transfer for non-parallel voice conversion by multi-scale style modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 3883–3895.

Yadav, A. K. S.; Bhagtani, K.; Xiang, Z.; Bestagini, P.; Tubaro, S.; and Delp, E. J. 2023. Dsvae: Interpretable disentangled representation for synthetic speech detection. *arXiv preprint arXiv:2304.03323*.

Yang, D.; Liu, S.; Huang, R.; Tian, J.; Weng, C.; and Zou, Y. 2023. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*.

Yao, J.; Yang, Y.; Lei, Y.; Ning, Z.; Hu, Y.; Pan, Y.; Yin, J.; Zhou, H.; Lu, H.; and Xie, L. 2024. Promptvc: Flexible stylistic voice conversion in latent space driven by natural language prompts. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 10571–10575. IEEE.

Yao, J.; Yuguang, Y.; Pan, Y.; Ning, Z.; Ye, J.; Zhou, H.; and Xie, L. 2025. Stablevc: Style controllable zero-shot voice conversion with conditional flow matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25669–25677.

Yuan, S.; Cheng, P.; Zhang, R.; Hao, W.; Gan, Z.; and Carin, L. 2021. Improving zero-shot voice style transfer via disentangled representation learning. *arXiv preprint arXiv:2103.09420*.

Zhang, J.-X.; Ling, Z.-H.; and Dai, L.-R. 2019. Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 540–552.

Zhang, M.; Wang, X.; Fang, F.; Li, H.; and Yamagishi, J. 2019. Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet. *arXiv preprint arXiv:1903.12389*.

Zhou, K.; Sisman, B.; Liu, R.; and Li, H. 2021. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 920–924. IEEE.