FC-TTS: Style and Timbre Control in Zero-Shot Text-to-Speech with Disentangled Speech Representations

Yoonhyung Lee and Hyunsin Park and Jinhwan Park and Jinkyu Lee*
Qualcomm AI Research†
{yoonhyun, hyunsinp, jinhpark, jinkyu}@qti.qualcomm.com

Abstract

Recent advances in zero-shot text-to-speech (TTS) have enabled accurate imitation of reference speech in terms of both speaking style and speaker timbre. However, achieving disentangled control over these aspects from separate references remains a challenging task. Several studies have proposed disentangled speech representations that decompose speech into interpretable attributes (*e.g.*, timbre, prosody, and content), providing a promising foundation for TTS with attribute control from separate references. Yet, how to effectively integrate such representations into TTS systems to achieve independent and precise control remains underexplored. In this paper, we present FC-TTS, a zero-shot TTS framework that enables disentangled control of style and timbre by conditioning on two distinct reference utterances. Unlike existing systems that inherit limitations from those pre-trained disentangled representations, FC-TTS introduces key design strategies, including architectural choices, training framework, and auxiliary training objectives, which improve the reliability of attribute separation and dual-reference control. Experiments show that FC-TTS achieves high-fidelity synthesis and competitive zero-shot naturalness, while uniquely supporting consistent and independent manipulation of style and timbre. Audio samples are available at https://qualcomm-ai-research.github.io/fc-tts

1 Introduction

With recent breakthroughs in text-to-speech (TTS) technology (Kim et al., 2021; Chen et al., 2025, 2024), the focus has expanded from simply generating natural-sounding speech to enabling expressive and personalized synthesis for diverse applications such as virtual assistants, audiobooks, accessibility tools, and interactive media (Xie et al., 2024). This evolution has brought a growing demand for fine-grained control over attributes such as speaking style and speaker timbre (Chen et al., 2025; Cho et al., 2025).

To support such control, existing TTS systems have explored various strategies. Supervised models trained on labeled datasets (Kim et al., 2021; Cho et al., 2024; Leng et al., 2024) offer reliable conditioning but scale poorly due to costly annotations. In contrast, reference-based zero-shot approaches (Casanova et al., 2022; Chen et al., 2025; Kim et al., 2023; Ji et al., 2024; Chen et al., 2024) offer greater flexibility by conditioning on example utterances, but typically entangle style and timbre in a single reference, hindering independent control.

To overcome this limitation, recent studies have explored disentangled speech representations (Choi et al., 2023; Ju et al., 2024), leading to various TTS systems that exploit such factorization. In particular, these systems employ disentangled attributes together with jointly pre-trained decoders to enable more interpretable control and, *in principle*, the possibility of conditioning on style and timbre separately. Nevertheless, disentanglement is often imperfect in practice, and the decoders inherently bound synthesis quality, with no guarantee of robustness to unseen style-timbre combinations.

In this paper, we investigate how to effectively leverage disentangled speech representations to build TTS models that support separate control over style and timbre using distinct references. To this end, we present FC-TTS, which leverages factorized speech representations but adds: (1) a two-stage spectrogram generation pipeline (timbre-conditioned blurry spectrogram, then prosody refinement) for robustness to unseen combinations; (2) a VQ-VAE-based style encoder that captures fine-grained and intra-utterance style variability;

*indicates the corresponding author.
†Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3772–3791

and (3) a conditioning-aware consistency loss that extends conventional regularization to multi-condition settings by enforcing joint coherence across timbre and style, providing more precise guidance for disentangled control.

In our experiments, we show that FC-TTS delivers competitive zero-shot TTS performance compared to state-of-the-art models that lack explicit support for separate control of style and timbre, as evaluated by objective metrics including UTMOS (Saeki et al., 2022), word error rate, and speaker similarity. Furthermore, we conduct detailed evaluations on the RAVDESS dataset (Livingstone and Russo, 2018a)—a highly expressive emotional speech corpus—focusing on both timbre and prosody controllability through both objective and subjective evaluations. Specifically, for timbre control, we compare FC-TTS with a factorized codec-based system (Ju et al., 2024), and for prosody control, we compare against a state-of-the-art zero-shot TTS model, F5-TTS (Chen et al., 2024). These evaluations highlight that FC-TTS not only maintains high synthesis quality but also enables precise and independent manipulation of style and timbre, which existing zero-shot systems do not explicitly support.

2 Related Work

2.1 Conditional Text-to-Speech

Conditional TTS incorporates auxiliary signals to control speaker or style attributes, enabling expressive and personalized speech. A common approach is to use labeled datasets with speaker or style annotations (Kim et al., 2021; Cho et al., 2024), which support multi-condition models capable of controlling multiple attributes simultaneously (Liu et al., 2023, 2025; Kang et al., 2023a). More recently, prompt-guided methods (Leng et al., 2024; Ji et al., 2025) have emerged, where models interpret free-form descriptions of style or tone. While these approaches offer disentangled controllability and flexibility, these approaches still rely heavily on annotation coverage and granularity, limiting scalability in open-domain or fine-grained control settings.

2.2 Disentangled Speech Representation Learning

A complementary direction focuses on learning disentangled speech representations that disentangle speech into interpretable components (*e.g.*, timbre, pitch, or linguistic content), opening promising directions for developing controllable TTS systems without labeled datasets. NANSY++ (Choi et al., 2023) leverages various information perturbation functions to isolate pitch, linguistic, and timbre features. FACodec (Ju et al., 2024) enforces strong information bottlenecks and uses supervision from phoneme and speaker labels to learn factorized representations. LSCodec (Guo et al., 2024) further improves timbre isolation via speaker perturbation and achieves strong results in voice conversion. Despite progress, most systems are trained in autoencoding setups with paired inputs and targets, meaning they are primarily optimized for naturally co-occurring factors. As a result, their ability to generalize to mismatched conditioning—such as combining timbre and prosody from different references—remains underexplored.

2.3 Reference Speech-based Style and Timbre Control

Recently, there have been studies to control the style and timbre of TTS models based on two reference samples. For instance, IndexTTS 2 (Zhou et al., 2025) achieves this by relying on the disentanglement property of semantic codec representations derived from MaskGCT (Wang et al., 2025). EmoSphere++ (Cho et al., 2025) trains a dedicated emotion encoder and introduces a regularization loss to enforce orthogonality between emotion and timbre vectors. Although promising, these methods rely on representations with only empirically observed disentanglement, or on pre-trained encoders that may discard other attributes such as accent. In contrast, our approach builds on systematically factorized representations that preserve reconstruction quality while enabling explicit decomposition, providing a more reliable basis for separate style and timbre control.

3 Methodology

3.1 Preliminaries

3.1.1 Factorized Speech Codec

To build a timbre and style controllable TTS system, we adopt FACodec[1] (Ju et al., 2024), a neural speech codec that decomposes speech into interpretable components while supporting faithful reconstruction. It factorizes a speech signal into multiple disentangled streams of discrete tokens, each capturing a distinct speech attribute: prosody

[1] https://github.com/open-mmlab/Amphion/tree/main/models/codec/ns3_codec

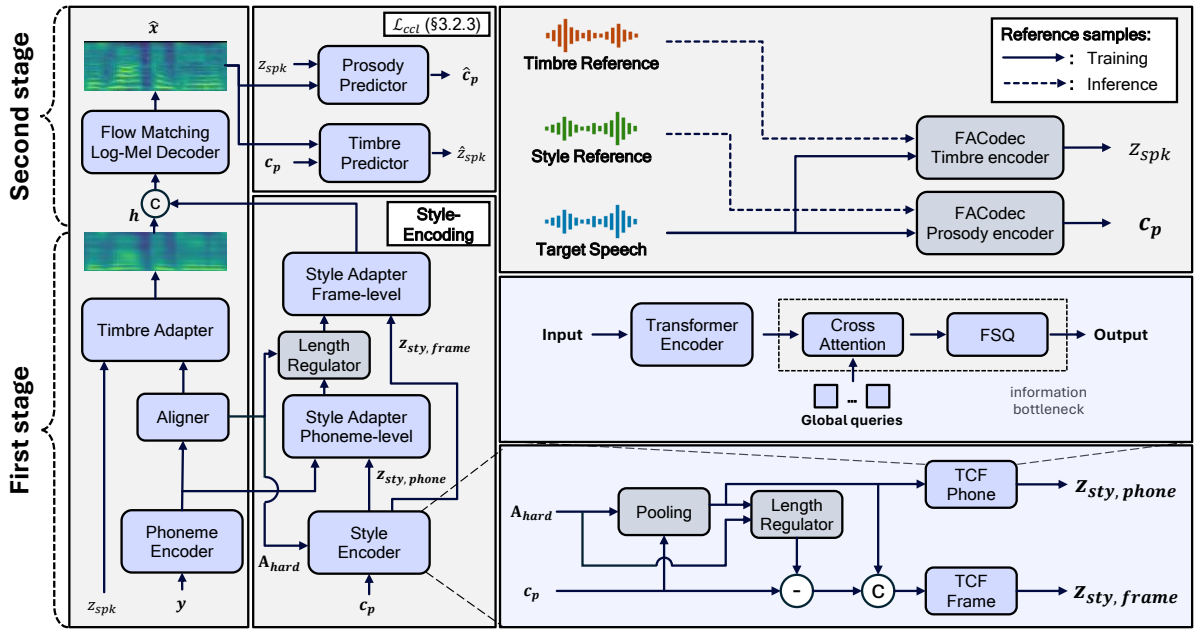


Figure 1: FC-TTS architecture. **First stage**: a phoneme sequence $\boldsymbol{y}$ is conditioned on timbre embedding z_{spk} via the timbre adapter to generate a blurry log-mel spectrogram $\boldsymbol{h}$, anchoring timbre characteristics. **Second stage**: $\boldsymbol{h}$ is refined into a clean spectrogram $\hat{\boldsymbol{x}}$ by a flow-matching decoder conditioned on style embedding $\boldsymbol{z}_{\text{sty}}$, obtained from prosody tokens $\mathbf{c_p}$ through hierarchical TCF modules, imprinting prosodic characteristics. During training, target speech provides both timbre and style references; at inference, separate references enable disentangled attribute control. Blue modules indicate trainable components.

tokens $\mathbf{c_p}$, content tokens $\mathbf{c_c}$, and acoustic detail tokens $\mathbf{c_d}$. Each stream is represented as $\mathbb{Z}^{N_* \times T}$, where T is the number of time steps with residual quantization levels $N_p = 1$, $N_c = 2$, and $N_d = 3$. In addition, speaker timbre is captured as a continuous global embedding $z_{\text{spk}} \in \mathbb{R}^D$. Notably, FC-TTS conditions exclusively on z_{spk} and $\mathbf{c_p}$; the content tokens $\mathbf{c_c}$ and acoustic detail tokens $\mathbf{c_d}$ are deliberately excluded to prevent information leakage that would compromise independent control of the two pathways.

3.1.2 Flow-matching TTS

In conditional TTS, the goal is to generate a target speech representation $\boldsymbol{x} \in \mathbb{R}^{F \times T}$ (a log-mel spectrogram in this work) given a phoneme sequence $\boldsymbol{y} \in \mathbb{Z}^L$ and additional conditioning information $\boldsymbol{c}$ such as speaker timbre or speaking style. To model this conditional generation process, we adopt the Conditional Flow-Matching (CFM) framework (Lipman et al., 2023), which defines a continuous-time transformation from a simple prior $p_1(\boldsymbol{x})$ (*e.g.*, isotropic Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$) to a target conditional distribution $p_0 = p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{c})$. To describe the progression from p_1 to p_0, CFM introduces a time-dependent flow $\phi_t \colon [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ that transports samples across time $t \in [0,1]$, with marginal distribution $p_t(\boldsymbol{x})$ at each step. This flow is driven by a velocity field $v_t(\boldsymbol{x}) \colon [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, which specifies the instantaneous direction of motion at each point. Their relationship is governed by the following ordinary differential equation (ODE):

$$\frac{d}{dt}\phi_t(\boldsymbol{x}) = v_t(\phi_t(\boldsymbol{x})), \quad \phi_1(\boldsymbol{x}) = \boldsymbol{x}_1. \tag{1}$$

Although the true v is not available in practice, CFM approximates it by training $u_\theta(\boldsymbol{x}, t, \boldsymbol{y}, \boldsymbol{c})$ with a conditional vector field $v_t(\boldsymbol{x}|\boldsymbol{x}_0)$. Among possible flows, the straight-line optimal transport (OT) trajectory is known to be efficient, and the corresponding ground-truth velocity is given by:

$$v_t^{\text{OT}}(\boldsymbol{x}_t|\boldsymbol{x}_0) = \boldsymbol{x}_0 - \boldsymbol{x}_1, \tag{2}$$

where $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_1 + t\boldsymbol{x}_0$. The model is then trained to align its predicted velocity u_θ with this OT velocity by minimizing the following loss:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{x}_1}\left[\|u_\theta(\boldsymbol{x}_t, t, \boldsymbol{y}, \boldsymbol{c}) - (\boldsymbol{x}_0 - \boldsymbol{x}_1)\|^2\right]. \tag{3}$$

3.2 FC-TTS

Figure 1 illustrates FC-TTS. Although the timbre condition z_{spk} and the style condition $\mathbf{c_p}$ are extracted from the same target during training, *in principle*, the factorized codec allows these attributes to be controlled separately at inference using references from different utterances. Building on this foundation, the following sections describe the architectural innovations that make such disentangled control practical. Concretely, FC-TTS processes the two conditions sequentially across dedicated stages: a *timbre stage* that anchors timbre characteristics via z_{spk} to produce a blurry spectrogram, followed by a *style stage* that imprints prosodic characteristics via $\mathbf{c_p}$ to refine it, ensuring each reference condition influences only its intended step. Additional implementation details are provided in Section 4.1 and Appendix A.

3.2.1 Hierarchical Spectrogram Generation

In preliminary experiments, we found that simply reusing the FACodec decoder, as in NaturalSpeech 3 (Ju et al., 2024), was insufficient for independent timbre-prosody control. The main reason is that the imperfect disentanglement provides no guarantee of robust generation on unseen combinations. To address this limitation, we propose a new model design that performs hierarchical log-mel spectrogram generation incorporating a jointly trained CFM speech decoder. It first generates a blurry spectrogram $\boldsymbol{h}$ using timbre information, and subsequently refines it into a complete spectrogram $\boldsymbol{x}_0$ using style information through the CFM decoder. These steps are trained jointly using mean-absolute-error (MAE) loss for the blurry spectrogram and CFM loss for the final output. The MAE objective, defined as $\mathcal{L}_{\text{blur}} = \mathbb{E}\left[\|\boldsymbol{h} - \boldsymbol{x}_0\|\right]$, encourages over-smoothed outputs—a property we exploit to avoid the need for pre-generated blurry spectrograms. Additionally, to prevent information leakage while maintaining consistent timbre and recording conditions, z_{spk} is randomly replaced with another utterance from the same long audio file. This two-stage design achieves functional separation: the timbre adapter injects z_{spk} in the first stage to anchor timbre characteristics, while the style adapter subsequently applies $\mathbf{c_p}$ to imprint prosodic characteristics, ensuring each reference influences a dedicated processing pathway.

3.2.2 VQ-VAE Style Encoding

Recent zero-shot TTS models perform well in mimicking voice characteristics based on in-context learning (ICL), where part of the target speech is used as a prompt and the model generates the rest, assuming consistent timbre and style. However, this assumption often fails in practice, as speaking style can vary even within a single utterance (Figure 2). To address this, we condition the model on style representations extracted from the target speech during training, eliminating the need to assume style consistency. However, this approach introduces a new challenge: the model may shortcut learning by copying surface-level acoustic features directly from the style reference, rather than capturing the intended higher-level prosodic patterns. For this purpose, we propose a style encoder module called TCF, combining a Transformer encoder, Cross-attention, and a Finite scalar quantization layer (FSQ), which is instantiated twice within the main architecture to hierarchically model style representations at both the phoneme and frame levels (Lei et al., 2023). Each design component of TCF specifically targets one of these challenges:

- **Prosody-only representation:** We exclusively use prosody tokens $\mathbf{c_p}$ from FACodec as the input to TCF, deliberately excluding content tokens $\mathbf{c_c}$ and acoustic detail tokens $\mathbf{c_d}$. This ensures that the style encoder captures rhythmic and intonational patterns without encoding unintended information.

- **Q-Former bottleneck (Li et al., 2023):** A fixed set of learned query tokens attends to the variable-length encoder outputs via cross-attention, compressing them into a fixed number of latent tokens. This bottleneck discards frame-level temporal details and forces the representation to retain only high-level stylistic structure, preventing the model from overfitting to the specific acoustic realization of the reference utterance.

- **Vector quantization (van den Oord et al., 2017):** The continuous latent tokens produced by the Q-Former are further discretized using FSQ (Mentzer et al., 2024). Quantization acts as an information bottleneck that suppresses low-level acoustic residuals and encourages the encoder to commit to a discrete, semantically meaningful style code.

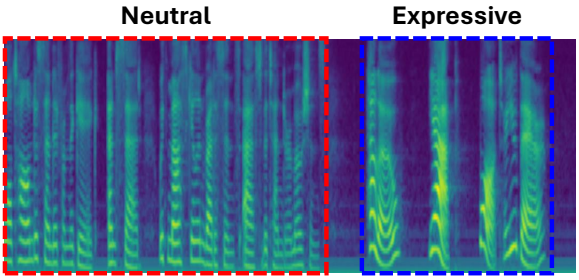


Figure 2: A speech sample from the Libriheavy dataset showing multiple speaking styles within a single utterance.

The detailed module architecture of TCF is provided in Appendix A.

3.2.3 Conditional Consistency Loss

To improve condition consistency in disentangled TTS, we present a conditional consistency loss (CCL) that extends prior regularization methods (Xin et al., 2021; Casanova et al., 2022) to multi-condition settings. By enforcing joint coherence across prosody and speaker identity, CCL provides more reliable guidance for disentangled control.

We first re-parameterize the CFM objective (Luo et al., 2025) so that the FC-TTS decoder generates log-mel spectrograms directly, instead of a vector field. Then, we train two attribute predictors on these spectrograms to predict the conditioning prosody token $\mathbf{c_p}$ and speaker embedding z_{spk} respectively. Importantly, each predictor also receives non-target conditioning signals, feeding z_{spk} to the prosody predictor and $\mathbf{c_p}$ to the timbre predictor. In Figure 3, the effect of this cross-conditioning is illustrated through an example scenario involving two conditional labels: gender and emotion.

The CCL is defined as a weighted sum of two terms: a cross-entropy loss for prosody prediction and a negative cosine similarity for speaker embedding consistency:

$$\mathcal{L}_{\text{CCL}} = \lambda_{\text{ccl-pro}} \cdot \mathbb{E}\left[\text{CE}(\mathbf{c_p}, f(\hat{\boldsymbol{x}}, z_{\text{spk}}))\right] - \lambda_{\text{ccl-spk}} \cdot \mathbb{E}\left[\cos\left(z_{\text{spk}}, g(\hat{\boldsymbol{x}}, \mathbf{c_p})\right)\right], \quad (4)$$

where $\text{CE}(\cdot,\cdot)$ denotes the cross-entropy loss, $f(\cdot)$ denotes the prosody predictor, and $g(\cdot)$ denotes the speaker embedding predictor.

4 Experiments

This section presents a thorough evaluation of FC-TTS, focusing on its zero-shot TTS performance and the disentanglement of style and timbre. We first describe the experimental setup, including datasets, model architecture, and training-inference procedures. We then compare FC-TTS with state-of-the-art (SOTA) baselines and conduct ablation studies to assess the contribution of individual components.

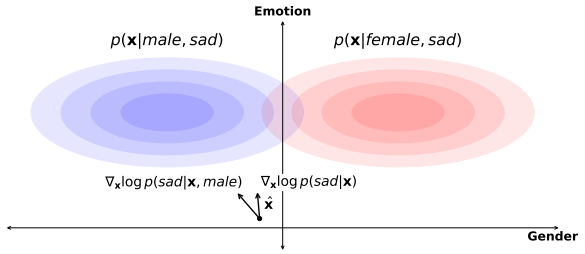


Figure 3: Gradients for CCL with multiple conditions. When a predictor takes only a log-mel spectrogram, the gradient $\nabla_{\boldsymbol{x}} \log p(\text{sad}|\boldsymbol{x})$ tends to point toward a region somewhere between the modes. In contrast, incorporating a known non-target attribute (*e.g.*, male) can sharpen the estimated posterior $p(\text{sad}|\boldsymbol{x}, \text{male})$, guiding spectrograms in more accurate directions, especially in early denoising steps when $\hat{\boldsymbol{x}}_0$ is only partially formed.

4.1 Experimental Setup

4.1.1 Datasets

For model training, we used the LibriHeavy dataset (Kang et al., 2023b), a large-scale audiobook corpus derived from LibriLight (Kahn et al., 2020). While it is not specifically curated for TTS, LibriHeavy offers substantial speaker and stylistic diversity, making it particularly useful for covering a wide range of speaking styles in zero-shot and expressive speech synthesis tasks. Our primary evaluation was conducted on the LibriSpeech `test-clean` subset, a widely adopted benchmark for zero-shot TTS. This split is especially suitable for testing generalization, as it contains speakers unseen during training. To further evaluate expressiveness and the disentanglement of style and timbre, we include experiments on the RAVDESS dataset (Livingstone and Russo, 2018b), which contains emotionally rich speech across multiple speakers. This dataset serves as a complementary test set for evaluating FC-TTS's ability to independently model and control timbre and style in complex, high-variance vocal contexts. Further details on datasets are provided in Appendix B.

4.1.2 Model Architecture

The phoneme encoder is implemented using a transformer encoder architecture. For alignment, we

adopt the RAD-TTS aligner (Shih et al., 2021) and train a CFM-based duration predictor using aligner-estimated durations. The timbre adapter is a transformer encoder where the speaker embedding z_{spk} conditions the layer normalization via adaptive layer normalization (Perez et al., 2018). The style adapter is a transformer decoder without causal masking, where style embeddings are provided through cross-attention layers. The log-mel spectrogram decoder consists of DiT blocks (Peebles and Xie, 2023). The TCF modules consist of a transformer encoder, a cross-attention layer with global queries, and a finite scalar quantization layer. Further architectural details are provided in Appendix A.

4.1.3 Training & Inference

We train FC-TTS on LibriHeavy for 200k iterations with AdamW (Loshchilov and Hutter, 2019), using a batch size of 64 and learning rate of 0.0002. During inference, FC-TTS first predicts durations using the duration predictor, with a number of function evaluations (NFEs) set to 8 and without classifier-free guidance. For log-mel spectrogram synthesis, we use 32 NFEs with a classifier-free guidance scale of 4.0, and apply random conditioning dropout with a probability of 15% during training. Finally, the generated log-mel spectrograms are converted to waveforms using HiFi-GAN (Kong et al., 2020). Further details on training and inference are provided in Appendix C.

4.1.4 Baselines

We benchmark FC-TTS against a diverse set of state-of-the-art TTS systems. Specifically, we include **NaturalSpeech 3** (Ju et al., 2024), which represents one of the strongest FACodec-based models, **F5-TTS** (Chen et al., 2024), which we retrain on LibriHeavy with a configuration comparable in size to ours, and reported results for **CLaM-TTS** (Kim et al., 2024) and **DiTTo-TTS** (Lee et al., 2025), representing recent advances in zero-shot speech synthesis. In addition, we introduce a **FACodec-based voice conversion (VC)** system, which serves as the upper bound of FACodec-based TTS performance. We adopt this setup for three reasons: (1) although several FACodec-based TTS models such as NaturalSpeech 3 have reported strong results, no official checkpoints are publicly available; (2) by directly reusing the ground-truth discrete tokens from the official FACodec encoder, this system approximates an idealized scenario in which a TTS model perfectly predicts codec tokens; and (3) by combining these tokens with an unmatched speaker embedding fed into the FACodec decoder, we simulate timbre controllability and assess the performance that a FACodec-based TTS system could achieve under such idealized conditions.

4.1.5 Metrics

We assess model performance using a comprehensive set of metrics: **UTMOS**[2] (Saeki et al., 2022) for speech quality, **WER**[3] for pronunciation accuracy, **SPK**[4] for speaker similarity, and **MCD** for prosodic and spectral similarity. These metrics collectively offer a multifaceted view of model performance across quality, intelligibility, and speaker consistency. Although MCD is not a perfect measure of prosodic similarity, we adopt it because it has been widely used in prior works as a practical proxy (Ju et al., 2024; Yang et al., 2025). Instead, we compute MCD with a fixed speaker to mitigate timbre confounding, so that the score primarily reflects prosodic differences. Additionally, we conduct a **human ABX listening test** focused on timbre and prosody, providing a more direct perceptual measure of controllability. To further supplement these evaluations with a more rigorous automated assessment of style similarity, we also employ an **AudioLLM-as-a-Judge** approach (Chiang et al., 2025; Manku et al., 2025) using Gemini 2.5 Pro (Comanici et al., 2025) as the evaluator. We measure two metrics: *Win Ratio* via an ABX test that determines which of two generated samples better matches the reference style, and *Style-MOS* that assigns a per-sample style similarity score (1–5). Further details on metric definitions and evaluation tools are described in Appendix D.

4.2 Results

4.2.1 Zero-Shot TTS Performance

Table 1 presents a comparative evaluation of FC-TTS against prior zero-shot TTS systems, focusing on naturalness and speaker similarity. Following common practice, we evaluate on the LibriSpeech `test-clean` split and report values from original papers when available. For evaluation, we use utterances from the `test-clean` subset that are 4 to 10 seconds long, with each text paired with a reference

[2] https://huggingface.co/spaces/sarulab-speech/UTMOS-demo

[3] https://huggingface.co/facebook/hubert-large-ls960-ft

[4] https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification

speech randomly sampled from the same speaker. While FC-TTS does not achieve the absolute best scores across all metrics, it demonstrates competitive performance relative to SOTA TTS systems. Importantly, unlike prior models, FC-TTS is explicitly designed for separate and controllable manipulation of timbre and style, a core capability not prioritized in existing TTS frameworks. Although NaturalSpeech 3 also supports reference-based control, its disentanglement ability is limited, as we discuss in the next section.

We attribute the performance gap to a combination of deliberate design choices in favor of stronger disentanglement and structural constraints introduced by our generation pipeline. First, we deliberately avoided using FACodec's content and detail tokens to prevent unintended information leakage; however, these tokens also carry useful cues that could enhance naturalness. In particular, detail tokens often encode recording-environment characteristics, and without them the model tends to converge toward the average acoustic condition of LibriHeavy, a corpus not originally optimized for TTS production quality. Second, our two-stage generation pipeline begins by producing a blurry spectrogram conditioned solely on timbre, which anchors the acoustic subspace that the subsequent style-based refinement stage must operate within. This structural constraint limits the distribution the flow-matching decoder can reach compared to end-to-end systems that jointly optimize all attributes without an intermediate bottleneck. While this design maximizes robustness to unseen style–timbre combinations, it may limit the ceiling of achievable naturalness relative to approaches such as NaturalSpeech 3 that impose no such constraint. Third, due to the imperfect disentanglement of FACodec, residual timbre cues may still reside in the excluded tokens; handling such leakage more precisely remains an open challenge. We view these as principled trade-offs that are necessary for achieving interpretable and independent controllability, and identify more robust disentanglement mechanisms—as well as codec-free alternatives—as key directions for future work.

4.2.2 Timbre Controllability on RAVDESS

In this experiment, we assess the model's robustness and its ability to accurately realize the intended timbre control when provided with prosodically richer reference samples from the RAVDESS dataset (Livingstone and Russo, 2018a). For each RAVDESS utterance used as a prosody reference, we synthesize speech by controlling the timbre according to two target speakers (one male and one female) selected from LibriSpeech `test-clean` subset.

	UTMOS ↑	WER ↓	SPK ↑	#Param. ↓
Ground-truth	4.10	2.07	0.71	-
HiFi-GAN	3.70	2.17	0.64	-
NaturalSpeech 3	**4.30**	**1.81**	**0.67**	500M
F5-TTS	-	2.42	0.66	336M
F5-TTS†	4.03	3.30	**0.67**	205M
DiTTo-TTS	-	2.69	0.60	508M
CLaM-TTS	-	5.11	0.50	584M
FC-TTS (ours)	4.22	1.88	0.60	204M

Table 1: Performance comparison of zero-shot TTS models. Reported values are taken from original papers if available. † indicates models re-trained on Libriheavy with similar model size.

	UTMOS ↑	SPK ↑	WER ↓	Win (%) ↑
FACodec-VC	3.19	0.27	8.40	10.7
Ours	**4.03**	**0.48**	**0.18**	**66.1**

Table 2: Objective metrics on the RAVDESS dataset for evaluating independent timbre control, a more challenging condition than standard TTS evaluation.

In this experiment, we use the **FACodec-based voice conversion (VC)** system as the comparison model. As described in Section 4.1.4, FACodec-VC directly reuses discrete tokens extracted from the FACodec encoder, approximating an oracle scenario where the codec tokens are perfectly predicted. Because there is no officially available FACodec-based TTS models (*e.g.*, NaturalSpeech 3), this configuration serves as a practical upper bound for the achievable performance of such systems. By pairing these ground-truth FACodec tokens with distinct speaker embeddings, FACodec-VC enables a controlled assessment of whether FC-TTS can maintain timbre controllability independently of prosodic variations.

As summarized in Table 2, FACodec-based timbre control exhibits substantial degradation across all objective metrics, achieving a UTMOS of 3.19, a SPK of 0.27, and a WER of 8.40. In contrast, FC-TTS maintains strong performance, reaching a UTMOS of 4.03, a SPK of 0.48, and a WER of 0.18[5], demonstrating its robustness under this

[5]The RAVDESS transcripts include only two short sentences — *"Kids are talking by the door."* and *"Dogs are sitting by the door."* — which explains the unusually low WER values.

	UTMOS ↑	SPK ↑	WER ↓	MCD ↓	Win (%) ↑
F5-TTS	3.40	**0.57**	4.39	3.43	8.9
Ours†	**3.95**	0.47	**0.30**	**3.21**	**65.5**

Table 3: Objective metrics on the RAVDESS dataset for evaluating independent style control. † indicates models using separate reference inputs for timbre and prosody.

	Win Ratio (%) ↑	Style-MOS ↑
F5-TTS	8.3	1.50
Ours	**91.7**	**3.92**

Table 4: AudioLLM-as-a-Judge (Gemini 2.5 Pro) evaluation for style control on RAVDESS.

mismatched setting. Furthermore, we conduct a subjective ABX test in which listeners are presented with (A) the FACodec-VC output, (B) the FC-TTS output, and (X) a target reference, and asked which of A or B is closer to X in terms of *timbre*. The resulting preference rate is reported as Win (%) in Table 2. Consistent with the objective results, FC-TTS is preferred in 66.1% of ABX trials, whereas FACodec-VC achieves only 10.7%. Taken together, these results demonstrate that FC-TTS enables effective and independent timbre control even in prosodically complex scenarios, reinforcing the practical advantage of our disentangled design beyond conventional codec-based TTS.

4.2.3 Prosody Controllability on RAVDESS

We next evaluate the prosody controllability of FC-TTS in comparison to F5-TTS using the RAVDESS dataset, as F5-TTS represents a SOTA zero-shot TTS system with publicly available source code. For a fair comparison, we retrain F5-TTS on the LibriHeavy dataset using phoneme inputs and match its model size to that of FC-TTS. Since F5-TTS does not support separate reference inputs for timbre and prosody, we provide a single reference speech during inference. In contrast, FC-TTS leverages two distinct references: an expressive RAVDESS utterance for prosody and a neutral utterance from the same speaker for timbre. This configuration imposes a more challenging inference condition but provides a clearer test of disentangled style control.

In addition to the automatic metrics used in the previous section, we further measure mel-cepstral distortion (MCD) between the expressive reference and the generated speech under matched text as an auxiliary indicator of prosody similarity. Although MCD is not a perfect proxy for prosodic attributes, we mitigate potential timbre confounding by fixing the timbre reference to a neutral utterance from the same speaker. We also complement these automatic evaluations with a subjective ABX prosody test, in which listeners judge which output—FC-TTS or F5-TTS—better matches the expressive reference in terms of prosodic pattern.

Results are summarized in Tables 3 and 4. FC-TTS outperforms F5-TTS in UTMOS (3.95 vs. 3.40), WER (0.30 vs. 4.39), and MCD (3.21 vs. 3.43), while also being preferred in 65.5% of ABX trials compared to only 8.9% for F5-TTS. These findings are further reinforced by the AudioLLM-as-a-Judge evaluation (Table 4): Gemini 2.5 Pro assigns FC-TTS a Win Ratio of 91.7% and a Style-MOS of 3.92, compared to only 8.3% and 1.50 for F5-TTS, respectively, providing strong evidence that the style differences perceived by human listeners are also reliably captured by an automated large-scale evaluator. Although FC-TTS yields a lower SPK score (0.47 vs. 0.57), this trade-off is consistent with the trends observed in Table 1; we note that maintaining perfect timbre consistency while achieving strong, disentangled style control from a separate reference—especially on highly expressive speech such as RAVDESS—represents an open research problem that we identify as a key direction for future work. Overall, these results collectively demonstrate that FC-TTS achieves more accurate and disentangled prosody transfer without sacrificing intelligibility, even under stricter evaluation conditions.

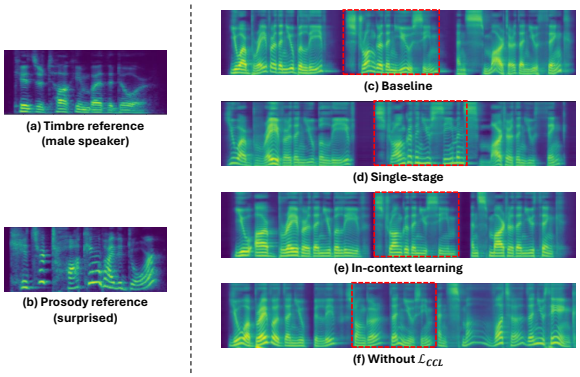


Figure 4: Log-mel spectrograms generated by each ablation variant (c-f), conditioned on the same timbre and prosody references (a, b) and using identical text input. The boxed regions highlight the segments corresponding to a specific phrase within the text input.

	Zero-shot TTS on LibriSpeech				Style Control on RAVDESS			
	UTMOS ↑	WER ↓	SPK ↑	MCD ↓	UTMOS ↑	WER ↓	SPK ↑	MCD ↓
FC-TTS	4.22	**1.88**	**0.60**	**5.60**	3.91	0.30	**0.37**	3.33
– two-stage generation	4.15	1.93	**0.60**	5.83	3.57	0.30	**0.37**	**3.26**
– VQ-VAE style encoding	**4.25**	2.00	0.57	5.62	**3.99**	**0.25**	0.34	3.47
– conditioning in consistency loss	4.21	1.92	0.59	5.67	3.79	0.35	0.36	3.36
– entire consistency loss	3.95	5.88	0.48	6.34	3.70	9.36	0.21	3.75

Table 5: Ablation study results. LibriSpeech use the same target data as reference for measuring MCD. RAVDESS uses the neutral speech as reference timbre and uses the emotional speech as reference style.

4.2.4 Ablation Study

We conduct an ablation study to analyze the contribution of three key components—two-stage generation, VQ-VAE style encoding, and conditional consistency loss—by systematically removing each from the full FC-TTS model. Table 5 summarizes quantitative results, and Figure 4 illustrates representative spectrograms that highlight the distinct role of each component in shaping prosody and timbre.

Two-stage generation. Removing this component degrades acoustic stability (UTMOS 4.15 vs. 4.22, MCD 5.83 vs. 5.60); Figure 4-(d) shows that the model over-reflects prosodic cues, yielding unstable acoustic patterns. The two-stage design decouples coarse timbre shaping from fine prosody refinement, improving robustness to unseen style–timbre combinations.

VQ-VAE style encoder. Removing this encoder yields a slight UTMOS gain (4.25 vs. 4.22) but notably weakens style control (SPK 0.57 vs. 0.60, MCD 3.47 vs. 3.33). As visible in Figure 4-(e), the generated speech exhibits flattened F0 contours that fail to follow the target prosody. Without the VQ-VAE encoder, the model falls back to ICL, which operates under an incorrect style assumption—treating the reference as stylistically uniform—and consequently fails to properly learn style reference following.

Consistency loss. Removing only the cross-conditioning moderately worsens prosody alignment and intelligibility (MCD 3.36 vs. 3.33; WER 1.92 vs. 1.88). Although the per-metric decrease is modest, this adds negligible training overhead and none at inference; the key design is that each predictor is conditioned on the other attribute (prosody predictor on z_{spk}, timbre predictor on $\mathbf{c_p}$), as visualized in Figure 3, which we believe is a useful building block for future multi-attribute control. Removing the *entire* consistency loss causes catastrophic degradation—WER 5.88 (LibriSpeech) and 9.36 (RAVDESS) vs. 1.88 and 0.30—with inconsistent pitch and rhythm visible in Figure 4-(f), confirming that conditional supervision is the most indispensable component overall.

In summary, both the quantitative metrics in Table 5 and the spectrograms in Figure 4 jointly demonstrate how each component contributes to prosody shaping and timbre stability. The results confirm that our proposed techniques—two-stage generation, VQ-VAE style encoding, and conditional consistency loss—play a central role in achieving robust and disentangled control. For a deeper perceptual understanding of these effects, we strongly recommend listening to the audio samples on our demo page.[6]

5 Conclusion

In this work, we presented FC-TTS, a zero-shot text-to-speech framework that enables independent control over timbre and speaking style using separate reference utterances. While built upon factorized codec representations, FC-TTS introduces three complementary innovations that extend beyond the codec's limitations: (1) a two-stage spectrogram generation pipeline improving robustness to unseen timbre-style combinations, (2) a VQ-VAE-based hierarchical style encoder that captures fine-grained intra-utterance variation, and (3) a conditional consistency loss enforcing coherence across conditioning factors. Extensive experiments on LibriSpeech and RAVDESS demonstrate that FC-TTS delivers competitive zero-shot quality and robust disentanglement, outperforming codec-based or single-reference baselines in both objective and perceptual evaluations. Overall, FC-TTS provides a solid foundation for future research in diverse expressive TTS applications requiring disentangled and controllable speech synthesis.

[6] https://qualcomm-ai-research.github.io/fc-tts

6 Limitations

Language coverage and generalization. Our training and evaluation are currently limited to English datasets, which under-represent many languages, dialects, and accents. This restriction may limit the model's ability to generalize to diverse linguistic conditions. Extending FC-TTS to multilingual and cross-accent scenarios will be an important step toward assessing its robustness and adaptability in broader expressive TTS settings.

Dependence on codec representations. Although FC-TTS introduces architectural and training innovations beyond codec-based systems, it still relies on the disentanglement quality of FACodec representations. Consequently, its absolute synthesis quality may appear slightly below the strongest SOTA TTS models. Future work could explore more robust disentanglement mechanisms or codec-free formulations to further enhance fidelity and controllability.

Definition and interpretability of attributes. The conceptual boundary between timbre and style remains an open question. For instance, determining whether a "husky voice" should be treated as a stylistic or timbral attribute is not trivial. Clarifying such definitions and establishing quantitative metrics that can reliably assess these dimensions would further advance the interpretability, controllability, and scientific rigor of expressive TTS research.

7 Ethical Considerations

The zero-shot capability of FC-TTS, while offering substantial flexibility for expressive and personalized synthesis, also introduces potential misuse risks, particularly in the creation of deepfake or impersonated speech. By enabling highly realistic synthesis from minimal reference data, such models can reproduce a speaker's unique timbre and convey new emotional or stylistic content without their consent, raising concerns around privacy, identity theft, and misinformation.

These risks become more pronounced as FC-TTS allows independent style control while preserving a fixed timbre, which can be exploited to generate emotionally manipulated speech in another person's voice. To mitigate such misuse, future deployments of expressive TTS systems might consider architectures or interfaces that restrict style controllability while keeping timbre generation confined to authorized sources. We hope that future research will further explore technical and policy-oriented safeguards to balance creative expression with ethical responsibility in controllable TTS technologies.

Acknowledgments

We are grateful to our colleagues for their support and encouragement throughout this research. In particular, we thank Guillaume Sautiere for insightful discussions and feedback on this work. The authors used AI writing assistance solely for minor language polishing and proofreading of the manuscript; it was not used to generate new content, ideas, or analyses.

References

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2709–2720. PMLR.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2025. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.

Cheng-Han Chiang, Xiaofei Wang, Chung-Ching Lin, Kevin Lin, Linjie Li, Radu Kopetz, Yao Qian, Zhendong Wang, Zhengyuan Yang, Hung-yi Lee, and Lijuan Wang. 2025. Audio-aware large language models as judges for speaking styles. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 467–480, Suzhou, China. Association for Computational Linguistics.

Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, Sang-Hoon Lee, and Seong-Whan Lee. 2024. Emosphere-tts: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech. In *Interspeech 2024*, pages 1810–1814.

Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, and Seong-Whan Lee. 2025. Emosphere++: Emotion-controllable zero-shot text-to-speech via emotion-adaptive spherical vector. *IEEE Transactions on Affective Computing*.

Hyeong-Seok Choi, Jinhyeok Yang, Juheon Lee, and Hyeongju Kim. 2023. NANSY++: Unified voice synthesis with neural analysis and synthesis. In *The Eleventh International Conference on Learning Representations*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Yiwei Guo, Zhihan Li, Chenpeng Du, Hankun Wang, Xie Chen, and Kai Yu. 2024. LSCodec: Low-bitrate and speaker-decoupled discrete speech codec. *arXiv preprint arXiv:2410.15764*.

Shengpeng Ji, Qian Chen, Wen Wang, Jialong Zuo, Minghui Fang, Ziyue Jiang, Hai Huang, Zehan Wang, Xize Cheng, Siqi Zheng, and Zhou Zhao. 2025. ControlSpeech: Towards simultaneous and independent zero-shot speaker cloning and zero-shot language style control. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6966–6981, Vienna, Austria. Association for Computational Linguistics.

Shengpeng Ji, Ziyue Jiang, Hanting Wang, Jialong Zuo, and Zhou Zhao. 2024. MobileSpeech: A fast and high-fidelity framework for mobile zero-shot text-to-speech. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13588–13600, Bangkok, Thailand. Association for Computational Linguistics.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Eric Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiangyang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In *ICML*.

J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. https://github.com/facebookresearch/libri-light.

Minki Kang, Wooseok Han, Sung Ju Hwang, and Eunho Yang. 2023a. Zet-speech: Zero-shot adaptive emotion-controllable text-to-speech synthesis with diffusion and style-based models. In *Interspeech 2023*, pages 4339–4343.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. 2023b. Libriheavy: a 50,000 hours asr corpus with punctuation casing and context. *Preprint*, arXiv:2309.08105.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In *Advances in Neural Information Processing Systems*, volume 33, pages 8067–8077. Curran Associates, Inc.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Jaehyeon Kim, Keon Lee, Seungjun Chung, and Jaewoong Cho. 2024. CLam-TTS: Improving neural codec language model for zero-shot text-to-speech. In *The Twelfth International Conference on Learning Representations*.

Sungwon Kim, Kevin J. Shih, Rohan Badlani, Joao Felipe Santos, Evelina Bakhturina, Mikyas T. Desta, Rafael Valle, Sungroh Yoon, and Bryan Catanzaro. 2023. P-flow: A fast and data-efficient zero-shot TTS through speech prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033.

Keon Lee, Dong Won Kim, Jaehyeon Kim, Seungjun Chung, and Jaewoong Cho. 2025. DiTTo-TTS: Diffusion transformers for scalable text-to-speech without domain-specific factors. In *The Thirteenth International Conference on Learning Representations*.

Shun Lei, Yixuan Zhou, Liyang Chen, Zhiyong Wu, Xixin Wu, Shiyin Kang, and Helen Meng. 2023. Msstyletts: Multi-scale style modeling with hierarchical context information for expressive speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3290–3303.

Yichong Leng, Zhifang Guo, Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yufei Liu, Dongchao Yang, leying zhang, Kaitao Song, Lei He, Xiangyang Li, sheng zhao, Tao Qin, and Jiang Bian. 2024. PromptTTS 2: Describing and generating voices with text prompt. In *The Twelfth International Conference on Learning Representations*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*.

Guanghou Liu, Yongmao Zhang, Yi Lei, Yunlin Chen, Rui Wang, Lei Xie, and Zhifei Li. 2023. Promptstyle: Controllable style transfer for text-to-speech with natural language descriptions. In *Interspeech 2023*, pages 4888–4892.

Jiaxuan Liu, Zhaoci Liu, Yajun Hu, Yingying Gao, Shilei Zhang, and Zhenhua Ling. 2025. Diff-StyleTTS: Diffusion-based hierarchical prosody modeling for text-to-speech with diverse and controllable styles. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5265–5272, Abu Dhabi, UAE. Association for Computational Linguistics.

Steven R. Livingstone and Frank A. Russo. 2018a. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):1–35.

Steven R Livingstone and Frank A Russo. 2018b. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Tianze Luo, Xingchen Miao, and Wenbo Duan. 2025. WaveFM: A high-fidelity and efficient vocoder based on flow matching. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2187–2198, Albuquerque, New Mexico. Association for Computational Linguistics.

Ruskin Raj Manku, Yuzhi Tang, Xingjian Shi, Mu Li, and Alex Smola. 2025. EmergentTTS-eval: Evaluating TTS models on complex prosodic, expressiveness, and linguistic challenges using model-as-a-judge. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. 2024. Finite scalar quantization: VQ-VAE made simple. In *The Twelfth International Conference on Learning Representations*.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. In *Interspeech 2022*, pages 4521–4525.

Kevin J Shih, Rafael Valle, Rohan Badlani, Adrian Lancucki, Wei Ping, and Bryan Catanzaro. 2021. Rad-tts: Parallel flow-based tts with robust alignment learning and diverse synthesis. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Neural Information Processing Systems*.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. 2025. MaskGCT: Zero-shot text-to-speech with masked generative codec transformer. In *The Thirteenth International Conference on Learning Representations*.

Tianxin Xie, Yan Rong, Pengfei Zhang, and Li Liu. 2024. Towards controllable speech synthesis in the era of large language models: A survey. *ArXiv*, abs/2412.06602.

Detai Xin, Yuki Saito, Shinnosuke Takamichi, Tomoki Koriyama, and Hiroshi Saruwatari. 2021. Cross-lingual speaker adaptation using domain adaptation and speaker consistency loss for text-to-speech synthesis. In *Interspeech 2021*, pages 1614–1618.

Yifan Yang, Bing Han, Hui Wang, Long Zhou, Wei Wang, Mingyu Cui, Xu Tan, and Xie Chen. 2025. Measuring prosody diversity in zero-shot tts: A new metric, benchmark, and exploration. *arXiv preprint arXiv:2509.19928*.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2025. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*.

A Architecture details

This section elaborates on the FC-TTS architecture, expanding on components that were just briefly described from the main paper due to space constraints. Detailed hyperparameter configurations are listed in Table 7.

- **FACodec encoder:** We utilize FACodec encoder to obtain disentangled speech representations: prosody tokens $\mathbf{c}_p$, content tokens $\mathbf{c}_c$, detail tokens $\mathbf{c}_d$, and the speaker embedding z_{spk}. Since the temporal resolution of the discrete tokens differs from that of the log-mel spectrograms, we upsample the tokens to match the spectrogram's frame rate using a repetition-based method. Additionally, we specifically utilize the tokens by first transforming them into embeddings via the codebook, followed by applying layer normalization without affine transformation to standardize the representations across their feature dimensions.

- **Text encoder:** The text encoder adopts a transformer-based architecture in which the traditional feed-forward network layers are substituted with one-dimensional convolutional layers. Positional information is encoded using rotary positional embeddings (Su et al., 2024). This configuration serves as the default transformer encoder architecture in this work.

- **Aligner:** This module aligns the output of the text encoder with the target log-mel spectrogram using an attention-based alignment search (Shih et al., 2021). To accelerate alignment training, it utilizes two loss functions: $\mathcal{L}_{\text{forwardsum}}$, which promotes monotonic diagonal alignment via the connectionist temporal classification (CTC) algorithm (Graves et al., 2006); and $\mathcal{L}_{\text{bin}}$, which encourages the soft alignment $\mathcal{A}_{\text{soft}}$ calculated from the attention mechanism to be close to the binarized hard alignment $\mathcal{A}_{\text{hard}}$ calculated from the monotonic alignment search algorithm (Kim et al., 2020). During the initial training phase, the aligner uses $\mathcal{A}_{\text{soft}}$ to expand the text representations and starts to use $\mathcal{A}_{\text{hard}}$ after 10,000 iterations to reduce the training-inference gap. Operations requiring alignment—such as pooling and length regulation—also use $\mathcal{A}_{\text{hard}}$.

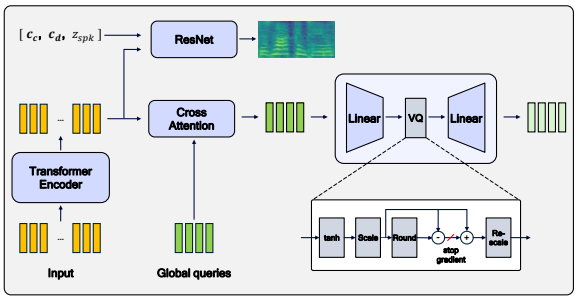


Figure 5: Detailed architecture of the TCF module. The input is first processed by a Transformer encoder, followed by compression using a Q-former-style cross-attention mechanism and a subsequence finite scalar quantization layer. To prevent representation collapse, a ResNet-based speech reconstruction module is jointly trained.

- **Style encoder:** Composed of two TCF modules operating at phoneme and frame levels respectively, this encoder models hierarchical style representations. It receives phoneme-level averaged prosody representations computed using $\mathcal{A}_{\text{hard}}$. For the frame-level input, it takes the concatenation of the re-expanded averaged prosody representations and the residual difference between the original and averaged representations to avoid encoding redundant information.

- **Timbre adapter:** It is built upon the transformer encoder with 1D CNNs. Also, the layer normalization is modified to adaptive layer normalization (AdaLN) (Perez et al., 2018), allowing for the injection of the speaker embedding vector z_{spk}.

- **Style adapter:** It is built upon the transformer decoder with 1D CNNs. Also, the style embedding encoded by the style encoder is fed to the style adapter via the cross attention layer. Here, we do not use the causal masking typically used in the transformer decoder.

- **Log-Mel decoder:** The decoder adopts the DiT architecture (Peebles and Xie, 2023), specifically the DiT Block with adaLN-Zero layers. Time embedding t is injected into the adaLN layers.

- **Prosody & Timbre predictors:** These attribute predictors are built using the transformer encoder architecture used in FACodec, allowing us to directly initialize them with

Training	
Learning Rate	0.0002
Batch Size	64
Iterations	200,000
Optimizer	AdamW
Weight Decay	0.0
Betas	(0.8, 0.99)
Warm-up	4,000
Decay type	exponential
Decay factor	0.999875
Decay step	200
Gradient clip norm	10.0

Table 6: Training hyperparameters

the transformer modules from the pre-trained FACodec encoder.

- **Duration predictor:** This module follows the duration predictor used in MaskGCT (Wang et al., 2025) and is trained using the CFM loss $\mathcal{L}_{\text{dur}}$ based on ICL. To encode duration references, it employs a transformer encoder block that processes a context prompt composed of log-scale durations, text representations, and phoneme-level averaged prosody representations extracted from a random segment corresponding to 0-30% of the target. It adopts DiT blocks incorporated with cross-attention layers to effectively integrate the duration prompt.

- **TCF module:** The TCF module comprises a transformer encoder, a cross-attention mechanism, and a finite scalar quantization (FSQ) layer (see Figure 5). During experimentation, we observed that the FSQ layer's latent representations tend to collapse into a single code. To mitigate this, we introduce an auxiliary ResNet module trained to reconstruct the log-mel spectrogram using a mean absolute error loss $\mathcal{L}_{\text{mel-recon}}$. The ResNet takes as input the output of the transformer encoder along with residual information necessary for spectrogram reconstruction.

B Dataset details

In this work, we use three datasets:

- **Libriheavy:** A large-scale English read speech corpus consisting of 50,000 hours of labeled audio derived from LibriVox and Librilight. It provides transcriptions with punctuation, casing, and textual context, enabling contextual ASR research. The dataset is divided into training subsets of 500h (small), 5,000h (medium), and 50,000h (large) and we use all these subsets for training. The dataset is available under an Apache-2.0 license at https://github.com/k2-fsa/libriheavy.

- **LibriSpeech:** A widely used 1,000-hour English ASR corpus based on public-domain LibriVox audiobooks and Project Gutenberg texts, distributed under the CC-BY 4.0 license. It includes training subsets of 100h, 360h, and 500h, and development and test sets of approximately 5 hours each (clean and other conditions). In this work, we exclusively used the *test-clean* subset (5.4 hours, 40 speakers: 20 male and 20 female) as our test set. From this subset, we selected only utterances between 4 and 10 seconds in duration, resulting in approximately 1.2k audio samples.

- **RAVDESS:** The Ryerson Audio-Visual Database of Emotional Speech and Song is a validated multimodal emotion dataset containing 7,356 recordings from 24 actors (12 male, 12 female). It includes audio-only, video-only, and audio-visual modalities across eight emotions for speech and six for song, each expressed at two intensity levels. The dataset is available under the CC BY-NC-SA 4.0 license.

C Training & Inference details

C.1 Training

FC-TTS is trained on LibriHeavy dataset for 200,000 iterations using AdamW (Loshchilov and Hutter, 2019) optimizer, with a batch size of 64 and beta parameters set to (0.8, 0.99), on 8 NVIDIA V100 GPUs for 116 hours. A learning rate of 0.0002 is employed, which undergoes a linear warmup over the initial 4,000 iterations before transitioning to exponential decay used in Kim et al. (2021). Also, gradient norms are clipped at a maximum of 10.0 to stabilize training.

The overall training objective is defined by the following composite loss function: $\mathcal{L}_{\text{total}} = \lambda_{\text{CFM}} \cdot \mathcal{L}_{\text{CFM}} + \lambda_{\text{blur}} \cdot \mathcal{L}_{\text{blur}} + \lambda_{\text{ccl-pro}} \cdot \mathcal{L}_{\text{CE}} + \lambda_{\text{ccl-spk}} \cdot \mathcal{L}_{\text{cossim}} + \lambda_{\text{mel-recon}} \cdot \mathcal{L}_{\text{mel-recon}} + \lambda_{\text{forwardsum}} \cdot \mathcal{L}_{\text{forwardsum}} + \alpha \cdot \lambda_{\text{bin}} \cdot \mathcal{L}_{\text{bin}} + \alpha \cdot \lambda_{\text{dur}} \cdot \mathcal{L}_{\text{dur}}$. Here, each λ_*

term controls the contribution of its corresponding loss component. The coefficients are set as follows: $\lambda_{\text{CFM}} = 5.0$, $\lambda_{\text{blur}} = 1.0$, $\lambda_{\text{ccl-pro}} = 0.2$, $\lambda_{\text{ccl-spk}} = 0.5$, $\lambda_{\text{mel-recon}} = 1.0$, $\lambda_{\text{dur}} = 1.0$, $\lambda_{\text{forwardsum}} = 0.1$, and $\lambda_{\text{bin}} = 0.1$. Additionally, the coefficient α is linearly warmed up over the first 10,000 iterations to allow time for the aligner to estimate accurate alignments.

C.2 Inference

Inference proceeds in two stages: duration prediction and spectrogram generation. Duration prediction is performed as the first stage of inference, using a fixed number of function evaluations (NFEs), set to 8. This stage does not incorporate classifier-free guidance. For log-mel spectrogram synthesis, we use 32 NFEs with a classifier-free guidance scale of 4.0. To enable classifier-free guidance, we randomly drop the conditioning inputs during training with a probability of 15%. Lastly, the generated log-mel spectrograms are transformed into 22kHz waveforms using a pre-trained HiFi-GAN (Kong et al., 2020) vocoder. To match the input requirements of HiFi-GAN, we use LibriHeavy speech samples that have been upsampled to 22kHz. Despite the waveform synthesis operating at 22kHz, the HiFi-GAN vocoder is trained with log-mel features configured with an `f_max` of 8000 Hz, which enables compatibility between features extracted from 16kHz audio and the final waveform synthesis at 22kHz. Further hyperparameter settings are listed in Table 6.

D Evaluation details

This section provides supplementary details on the evaluation setup and methodology. We compare the performance of FC-TTS against a diverse set of baseline models to ensure a comprehensive evaluation:

D.1 Baselines

- **NaturalSpeech 3 (NS3)**: Selected as the primary baseline due to its strong performance and architectural similarity to FC-TTS, notably its use of FACodec. When we asses its separate controllability of timbre and style, we instead evaluate the voice conversion capability of FACodec, which serves as an upper bound for NS3's performance in this context.

- **F5-TTS** (Chen et al., 2024): An in-context learning-based state-of-the-art zero-shot TTS model. We selected this model because its official source code is publicly available under MIT License, allowing us to retrain it under identical conditions and thereby ensure both fairness in evaluation.

- **CLaM-TTS & DiTTo-TTS** (Kim et al., 2024; Lee et al., 2025): These models are also included to broaden the comparison, using reported metrics evaluated on the same test set and scoring protocol.

- **FACodec** (Ju et al., 2024): This is used a voice conversion system to approximate the oracle of the FACodec-based TTS models. Based on the official checkpoint, which is available under MIT License, we first extract ground-truth FACodec tokens from the encoder and reconstruct it with unmatched timbre embedding using FACodec decoder.

D.2 Metrics

Also, we assess model performance using a diverse set of objective metrics as follows:

- **UTMOS**[7] (Saeki et al., 2022) (MIT License): A neural network-based metric trained to predict human mean opinion scores (MOS) for speech quality. It provides a reliable proxy for perceptual naturalness and fluency, and is widely adopted for evaluating TTS systems.

- **WER (Word Error Rate)**: Computed using a pre-trained HuBERT-based automatic speech recognition model[8] (Apache 2.0 License). WER quantifies intelligibility by comparing the transcribed output of generated speech against ground-truth text, with lower values indicating better pronunciation accuracy.

- **SPK**: A metric for evaluating speaker similarity, computed as the cosine similarity between embeddings extracted using a WavLM-TDCNN-based speaker verification model[9] (CC BY-SA 3.0 License). These embeddings are obtained from both reference and synthesized speech samples. Higher scores indicate greater similarity in speaker characteristics.

[7] https://huggingface.co/spaces/sarulab-speech/UTMOS-demo

[8] https://huggingface.co/facebook/hubert-large-ls960-ft

[9] https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification

- **MCD (Mel Cepstral Distortion)**: MCD calculates the Euclidean distance between mel-cepstral coefficients of two speech signals. These coefficients capture the spectral envelope of speech. In our evaluation, we use MCD for assessing prosodic similarity between the reference and generated utterances, which contain identical linguistic content and are spoken by the same speaker.

D.3 ABX test

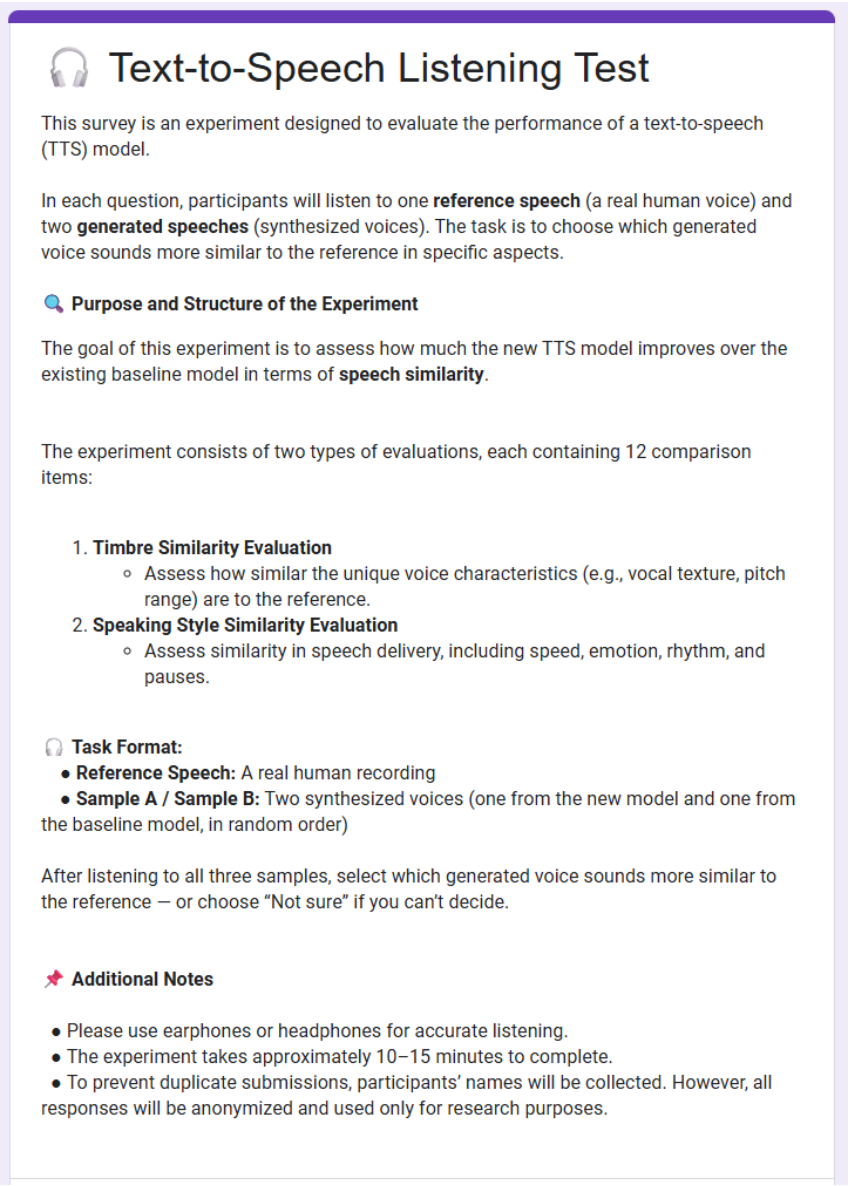
The human listening evaluation was conducted with internal company employees serving as participants. To ensure fairness and reliability, a set of test samples was randomly selected from the evaluation dataset. For each sample, corresponding speech outputs were generated by both the baseline system and our proposed system. During the test, participants were presented with pairs of audio clips in randomized order—one from each system—without any indication of their origin. They were asked to identify which sample more closely matched the reference in terms of naturalness and intelligibility. This procedure was designed to minimize potential bias and to ensure the validity and reproducibility of the evaluation results.

D.4 AudioLLM-as-a-Judge Evaluation

We use Gemini 2.5 Pro (Comanici et al., 2025) as the evaluator model and measure two complementary metrics. **Win Ratio** is obtained via pairwise comparison: the model is given a human reference and two generated samples and judges which better matches the reference speaking style; the aggregate preference rate over all test pairs constitutes the Win Ratio. **Style-MOS** is obtained via independent per-sample scoring: the model rates how well a single generated sample reflects the speaking style of a given reference on a 1–5 scale, and the mean over all pairs constitutes the Style-MOS.

Module	Hyperparameter	Value
Text encoder / Duration prompt encoder	Hidden dimension	192
	Layers	6
	Heads	2
	FFN dimension	768
	Conv kernel size	3
Duration predictor	Hidden dimension	192
	Layers	6
	Heads	4
	FFN dimension	768
	Conv kernel size	3
Timbre adapter / Style adapter-phone / Style adapter-frame	Hidden dimension	192
	Layers	4
	Heads	4
	FFN dimension	768
	Conv kernel size	9 / 5 / 9
Log-mel decoder	Hidden dimension	384
	Layers	12
	Heads	4
	FFN dimension	768
	Conv kernel size	5
Prosody / Timbre predictor	Hidden dimension	256
	Layers	4
	Heads	4
	FFN dimension	1024
	Conv kernel size	5
TCF-phone / TCF-frame	Hidden dimension	384
	Layers	6
	Heads	4
	Conv kernel size	5
	Queries	4
	FSQ latent dimension	6
	FSQ latent bins	[5, 5, 5, 5, 5, 5]
	ResNet Blocks	3
	ResNet Conv Layers	2
	ResNet Conv kernel size	[[9, 1], [9, 1], [9, 1]]
	ResNet Conv dilations	[[1, 1], [2, 1], [4, 1]]

Table 7: Grouped hyperparameters for different modules. Modules with similar configurations are merged into single group, and differing values are separated by slashes (e.g., "Timbre adapter / Style adapter-phone / Style adapter-frame") to reduce redundancy and improve readability.



🎧 Text-to-Speech Listening Test

This survey is an experiment designed to evaluate the performance of a text-to-speech (TTS) model.

In each question, participants will listen to one **reference speech** (a real human voice) and two **generated speeches** (synthesized voices). The task is to choose which generated voice sounds more similar to the reference in specific aspects.

🔍 **Purpose and Structure of the Experiment**

The goal of this experiment is to assess how much the new TTS model improves over the existing baseline model in terms of **speech similarity**.

The experiment consists of two types of evaluations, each containing 12 comparison items:

1. **Timbre Similarity Evaluation**
   - Assess how similar the unique voice characteristics (e.g., vocal texture, pitch range) are to the reference.
2. **Speaking Style Similarity Evaluation**
   - Assess similarity in speech delivery, including speed, emotion, rhythm, and pauses.

🎧 **Task Format:**

- **Reference Speech:** A real human recording
- **Sample A / Sample B:** Two synthesized voices (one from the new model and one from the baseline model, in random order)

After listening to all three samples, select which generated voice sounds more similar to the reference — or choose "Not sure" if you can't decide.

📌 **Additional Notes**

- Please use earphones or headphones for accurate listening.
- The experiment takes approximately 10–15 minutes to complete.
- To prevent duplicate submissions, participants' names will be collected. However, all responses will be anonymized and used only for research purposes.

Figure 6: Main instructions provided to participants in the ABX test.

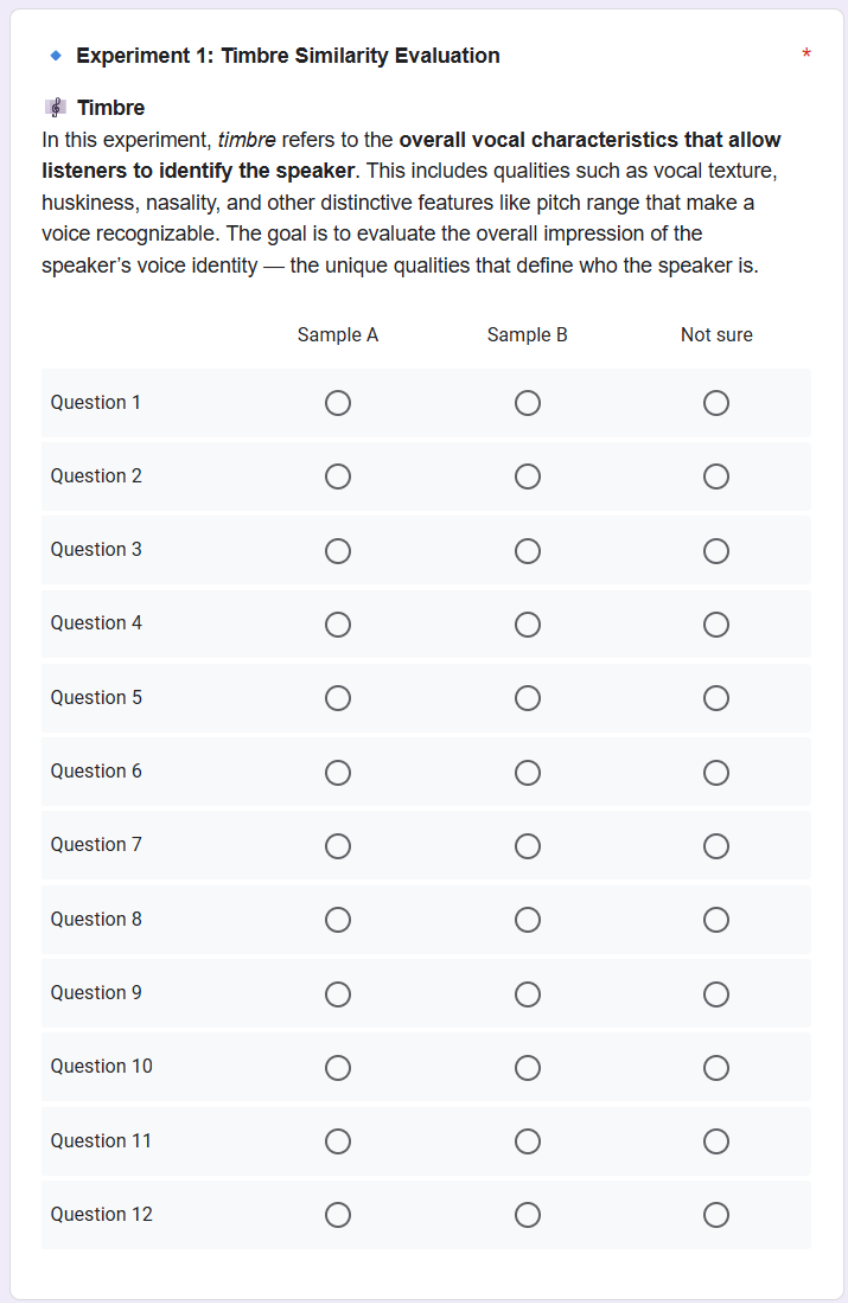


Figure 7: Instructions for the timbre controllability evaluation.

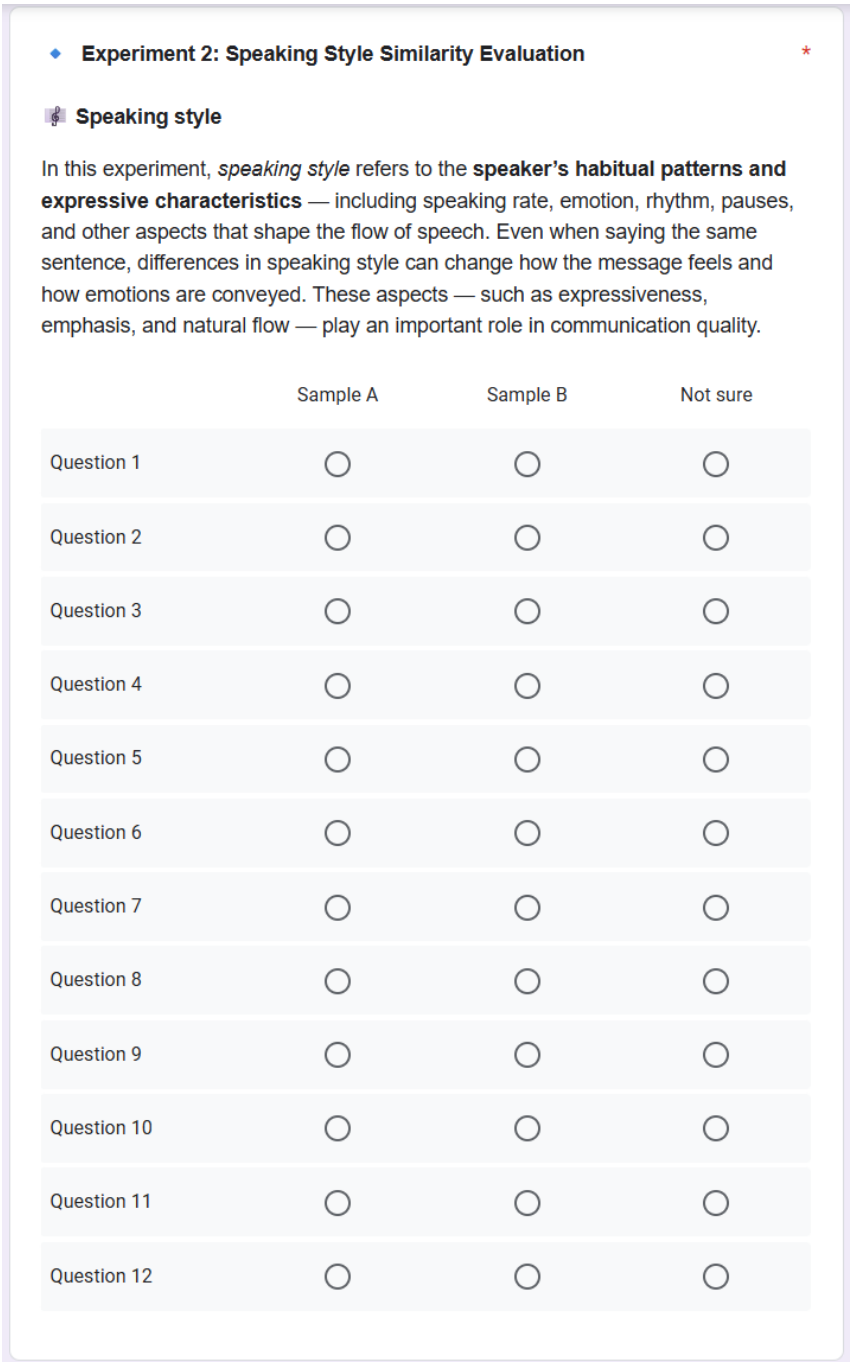


Figure 8: Instructions for the style controllability evaluation.