CoCiTTS: Towards Character-Level Speaker Consistency in Zero-Shot TTS for Cross-Lingual Cinematic Dubbing

Anonymous authors
Paper under double-blind review

Abstract

Cross-Lingual Cinematic Dubbing aims to synthesize target-language speech from source-language speech in movies or TV shows, which requires a character to retain a consistent voice across scenes with different emotional expressions. However, existing zero-shot text-to-speech (TTS) systems mainly optimize utterance-level speaker similarity and overlook character-level speaker consistency across multiple utterances. This limitation becomes particularly challenging in cinematic scenarios, where the same character frequently exhibits diverse emotional states, and emotion-rich reference utterances may lead to inconsistent speaker timbre across generated utterances. In this paper, we identify emotion leakage within speaker representations as an important factor contributing to speaker timbre drift. To address this issue, we propose **CoCiTTS** (**Co**nsistency-Oriented **Ci**nematic Text-to-Speech), a zero-shot TTS framework that learns consistency-oriented speaker conditioning for cross-lingual cinematic dubbing. CoCiTTS introduces a Speaker Timbre Purification (STP) module based on adversarial disentanglement to suppress emotion leakage while preserving speaker identity. Furthermore, a Fine-grained Emotion Compensation (FEC) module models global and local emotional characteristics through an independent pathway, enabling expressive generation without contaminating speaker representations. In addition, we construct **CineSpeech**, an emotion-rich cinematic speech benchmark containing 136 characters, and establish consistency-aware evaluation methods for character-level timbre preservation. Extensive experiments on multilingual cross-lingual tasks demonstrate that CoCiTTS effectively improves speaker consistency while maintaining competitive speech quality, speaker similarity, and emotional expressiveness compared with existing zero-shot TTS systems.

1 Introduction

Cross-Lingual Cinematic Dubbing is a task of synthesizing target-language speech from source-language cinematic speech while preserving characters' identities and expressive delivery. As cultural content rapidly globalizes, automatic dubbing has drawn increasing attention as an efficient alternative to traditional manual production. Recently, zero-shot text-to-speech (TTS) models, XTTS (Casanova et al., 2024), the CosyVoice series (Du et al., 2024a;b; 2025), IndexTTS2 (Zhou et al., 2026), and Qwen3-TTS (Hu et al., 2026), have achieved remarkable progress in multilingual speech synthesis by extracting characteristics for an unseen speaker from a short reference utterance. These advances provide a promising solution for scalable cinematic dubbing systems.

However, existing zero-shot TTS systems primarily focus on utterance-level speaker similarity (SS), while overlooking character-level speaker consistency. We observe that generated utterances from the same character can exhibit high SS with their references yet still sound different from one another. This distinction matters in cinematic dubbing, where the soundtracks of a character usually contain rich and diverse emotional expressions like happiness, sadness, and anger (Figure 1). Therefore, a practical cinematic dubbing system should not only reproduce the speaker characteristics of each reference utterance but also maintain a stable and consistent timbre throughout the entire movie. However, this consistency requirement remains largely unexplored in existing zero-shot TTS approaches.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Recently, several works have explored automatic movie dubbing (Chen et al., 2022), including InstructDubber (Zhang et al., 2026) and FunCineForge (Liu et al., 2026). Specifically, given a silent video clip, dubbing script, and reference speech, these methods leverage multimodal large language models (MLLMs) to infer speaking styles and emotional states, and to generate visually aligned dubbing speech. In contrast, our work focuses on cross-lingual zero-shot TTS with existing original film audio. Our goal is to synthesize target-language speech that preserves character identity, timbre, and emotional expression of the original cinematic soundtracks. Unlike the above methods for movie dubbing, which generate speech conditioned on silent video clips, our work aims to transfer existing film and television content across languages using zero-shot TTS technology.

In this paper, we formulate character-level speaker consistency under emotional variations as a distinct evaluation and modeling problem for cross-lingual cinematic dubbing. We identify that emotion leakage in speaker representations is an important factor contributing to speaker inconsistency. As shown in Figure 1, speaker representations extracted from expressive reference utterances may encode not only identity but also emotional information. When the TTS model is conditioned on such representations, emotionally different reference speeches may lead to inconsistent speaker timbre in generated utterances of the same character. Furthermore, an emotion classifier achieves an accuracy over 80% on these speaker representations, indicating that emotion-related information is encoded in speaker representations (see the Appendix A.2.1 for details). These observations motivate the need for learning consistency-oriented speaker conditioning that preserves speaker identity while reducing undesirable emotion-related variations.

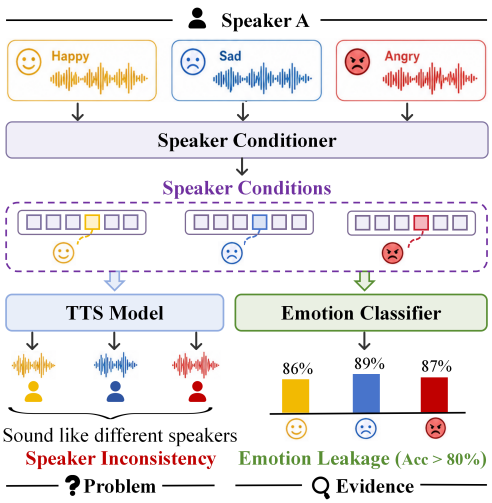


Figure 1: Motivation of CoCiTTS. Problem: The generated utterances from emotion-rich cinematic references sound like different speakers. Evidence: Speaker conditions achieve over 80% emotion classification accuracy.

Based on this insight, we propose **CoCiTTS** (**Co**nsistency-Oriented **Ci**nematic Text-to-Speech), a zero-shot TTS framework designed for cross-lingual cinematic dubbing. Rather than redesigning the entire TTS architecture, CoCiTTS focuses on optimizing the speaker conditioning mechanism and improving character-level speaker consistency. Specifically, CoCiTTS uses a Speaker Timbre Purification (STP) module based on adversarial disentanglement to suppress emotion leakage in speaker conditioning while retaining speaker identity. Furthermore, a Fine-grained Emotion Compensation (FEC) module is introduced to model emotional information through an independent global-local pathway. This design targets a specific trade-off exposed by cinematic dubbing: stabilizing character timbre across emotionally diverse references without suppressing expressive delivery. To systematically evaluate this problem, we construct CineSpeech, an emotion-rich cinematic speech benchmark containing 950 utterances from 136 cinematic characters. Unlike existing speech datasets, CineSpeech is designed to assess character-level speaker consistency under diverse emotional conditions. Based on CineSpeech, we establish multilingual cross-lingual dubbing tasks and introduce consistency-aware evaluation protocols. Extensive experiments on CineSpeech and the open-source CineDub-CN (Liu et al., 2026) dataset demonstrate that CoCiTTS significantly improves speaker timbre consistency compared with existing zero-shot TTS systems while maintaining competitive WER, UTMOS, speaker similarity, and emotion similarity.

The contributions of this paper are summarized as follows:

1) We identify and formulate character-level speaker consistency under emotionally diverse references as a distinct challenge for cross-lingual cinematic dubbing.

2) We systematically analyze the relationship between emotional variation and speaker timbre drift, revealing that emotion leakage within speaker conditioning contributes to speaker timbre drift and cross-emotion references are associated with consistency degradation.

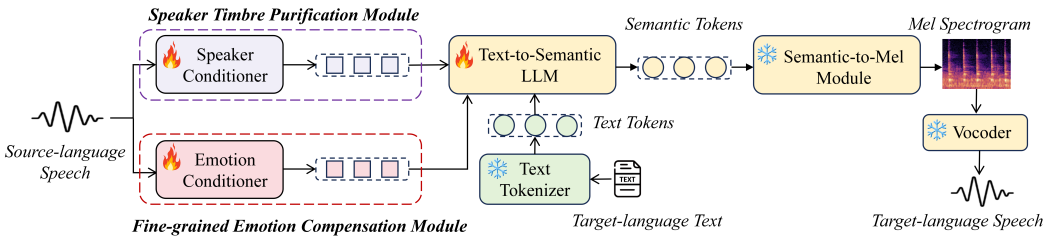


Figure 2: Overall architecture of CoCiTTS.

3) We propose CoCiTTS, a consistency-oriented TTS framework that decouples speaker and emotion modeling through STP and FEC modules. Experiments demonstrate that CoCiTTS effectively improves speaker consistency while maintaining speech quality and emotional expressiveness.

4) We construct CineSpeech, an emotion-rich cinematic speech benchmark, together with consistency metrics and multilingual cross-lingual tasks, providing a systematic testbed for studying character-level speaker consistency beyond utterance-level similarity.

2 Related Work

Zero-shot TTS. Early TTS works incorporated discrete speech representations and diffusion models, such as VALL-E (Chen et al., 2025), NaturalSpeech (Shen et al., 2024; Ju et al., 2024), and Mega-TTS (Jiang et al., 2024; 2025). Recent advances in zero-shot TTS have enabled high-quality speech synthesis for unseen speakers using only a few seconds of reference speech. Recent LLM-based systems, including XTTS (Casanova et al., 2024), CosyVoice series (Du et al., 2024a;b; 2025), IndexTTS2 (Zhou et al., 2026), and Qwen3-TTS (Hu et al., 2026), further improve multilingual synthesis through large-scale pretraining and semantic speech modeling. However, existing zero-shot TTS systems mainly optimize utterance-level speaker similarity while paying limited attention to character-level speaker consistency, which is a key requirement for cross-lingual cinematic dubbing.

Movie Dubbing. Movie dubbing methods such as InstructDubber (Zhang et al., 2026) and FunCineForge (Liu et al., 2026) aim to perform video-conditioned speech generation, where a silent video clip, a dubbing script, and a reference speech are jointly modeled to synthesize speech. In contrast, our work studies a different setting: cross-lingual transfer of existing cinematic speeches, where source-language cinematic utterances are converted into target-language utterances while preserving speaker identity and emotional expression.

3 Method

As shown in Figure 2, CoCiTTS is a cascaded autoregressive TTS model built upon IndexTTS2 (Zhou et al., 2026) with two key modifications. It aims to improve character-level speaker consistency in cross-lingual cinematic dubbing by learning consistency-oriented speaker conditioning. In this section, we describe the problem formulation, the Speaker Timbre Purification module, the Fine-grained Emotion Compensation module, and the training objective.

3.1 Problem Formulation

In cross-lingual cinematic dubbing, given a speaker p and N reference utterances that may exhibit different emotions, the reference set is defined as

$$\mathcal{X}_p = \left\{ \mathbf{x}_p^{(1)}, \mathbf{x}_p^{(2)}, \ldots, \mathbf{x}_p^{(N)} \right\}, \tag{1}$$

$$\exists\, i \neq j \quad \text{s.t.} \quad y_p^{(i)} \neq y_p^{(j)}, \tag{2}$$

where $\mathbf{x}_p^{(i)}$ denotes the i-th reference utterance of speaker p and $y_p^{(i)}$ denotes its emotion label. The original speaker representation $\mathbf{s}_p^{(i)}$ is obtained by the speaker conditioner $E_{spk}(\cdot)$ of the backbone

model. The emotion representation $\mathbf{e}_p^{(i)}$ is obtained by the emotion conditioner $E_{emo}(\cdot)$:

$$\mathbf{s}_p^{(i)} = E_{spk}\left(\mathbf{x}_p^{(i)}\right), \mathbf{e}_p^{(i)} = E_{emo}\left(\mathbf{x}_p^{(i)}\right). \tag{3}$$

However, as illustrated in Figure 1, speaker representations extracted from emotionally expressive reference speeches inevitably contain emotion-related information in cinematic scenarios. Conceptually, the extracted speaker representation can be regarded as an entangled function of a speaker-related factor and an emotion-related factor:

$$\mathbf{s}_p^{(i)} = f\left(\mathbf{z}_p^{\text{spk}}, \mathbf{z}_{p,i}^{\text{emo}}\right), \tag{4}$$

where $\mathbf{z}_p^{\text{spk}}$ represents speaker identity, $\mathbf{z}_{p,i}^{\text{emo}}$ represents emotion-related variation, and $f(\cdot)$ denotes their entanglement in the original speaker representation, leading to speaker timbre drift when reference speeches exhibit diverse and strong emotional states.

3.1.1 Speaker Timbre Purification Module

To obtain emotion-invariant speaker conditioning, we propose the Speaker Timbre Purification (STP) module, as illustrated in Figure 3 (a). We adopt an adversarial representation learning framework to disentangle timbre and emotion. Specifically, the speaker representation $\mathbf{s}$ output from the speaker conditioner $E_{spk}(\cdot)$ is fed into two branches at the same time. The first branch is an emotion classifier preceded by a Gradient Reversal Layer (GRL). During backpropagation, GRL_λ multiplies the gradient propagated to the speaker conditioner by $-\lambda$, while the emotion classifier is optimized using the standard cross-entropy gradient. Consequently, the speaker conditioner learns to generate features that are difficult for the emotion classifier to classify correctly, while preserving speaker identity.

$$\mathcal{L}_{\text{emo_adv}} = \text{CE}\left(C_{emo}\left(\text{GRL}_\lambda(\mathbf{s})\right), y\right), \tag{5}$$

where $C_{emo}(\cdot)$ denotes the emotion classifier, and y denotes the ground-truth emotion label.

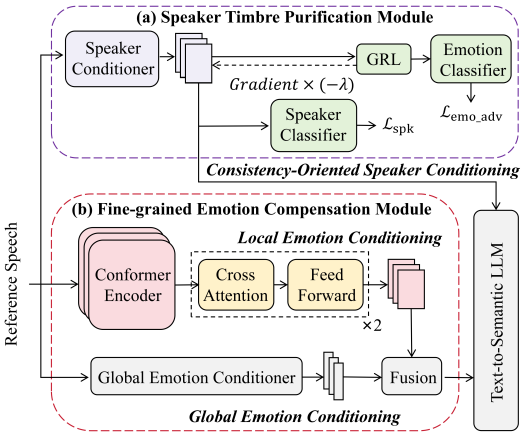


Figure 3: Details of the proposed framework: (a) Speaker Timbre Purification Module and (b) Fine-grained Emotion Compensation Module.

The second branch is a speaker classifier to ensure that the speaker conditioner can accurately extract identity and timbre information. We adopt ArcFace loss (Deng et al., 2019) as the speaker classification objective to enhance the discriminability of speaker representations while preserving speaker identity,

$$\mathcal{L}_{\text{spk}} = \mathcal{L}_{\text{ArcFace}}\left(C_{spk}(\mathbf{s}), p\right), \tag{6}$$

where $C_{spk}(\cdot)$ denotes the speaker classifier, and p denotes the ground-truth speaker identity. The overall objective of the STP module is

$$\mathcal{L}_{\text{STP}} = \mathcal{L}_{\text{spk}} + \mathcal{L}_{\text{emo_adv}}. \tag{7}$$

Through the two-branch architecture, the speaker conditioner is optimized to learn timbre-relevant speaker representations and reduce emotion leakage, thereby providing consistency-oriented conditioning for the TTS backbone.

3.2 Fine-grained Emotion Compensation Module

The Speaker Timbre Purification module inevitably suppresses emotional cues that are beneficial for expressive speech synthesis. Therefore, to ensure that the generated speech maintains rich and natural emotional expression, an independent emotion pathway is introduced to compensate for the removed emotional information while avoiding emotion-contaminated speaker conditioning. Specifically, we introduce a Fine-grained Emotion Compensation (FEC) module to model global and local

emotion conditioning, as shown in Figure 3 (b). We directly adopt the emotion conditioner from the backbone model to obtain the global emotion conditioning:

$$\mathbf{e}_{global} = E_{global}(\mathbf{x}), \tag{8}$$

which can describe the overall emotional style of an utterance but fails to sufficiently model the fine-grained emotional variations within the speech.

To this end, we introduce a local emotion branch. Specifically, the reference speech $\mathbf{x}$ is first encoded by a local emotion encoder $E_{local}(\cdot)$ to obtain frame-level emotion features H_{emo}:

$$H_{emo} = E_{local}(\mathbf{x}). \tag{9}$$

Then we use a set of learnable latent queries Q to attend to the frame-level emotion features through cross-attention:

$$\mathbf{e}_{local} = CrossAttn\left(Q, H_{emo}, H_{emo}\right), \tag{10}$$

where $Q \in \mathbb{R}^{32\times d}$, d denotes the feature dimension and is set to 1280. $\mathbf{e}_{local}$ denotes the local emotion conditioning. Q, H_{emo} and H_{emo} serve as the query, key, and value of the cross-attention operation, respectively.

3.3 Training Objective

To effectively disentangle speaker identity from emotional variations and avoid interference between speaker purification and emotion compensation, we adopt a two-stage training strategy. In the first stage, we optimize the Speaker Timbre Purification (STP) module together with the text-to-semantic LLM module, while all remaining modules are kept frozen during the first stage. This stage aims to reduce emotion leakage while preserving the speaker identity. The training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{mel}} + \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{STP}}, \tag{11}$$

where L_{mel} and L_{text} measure the precision of discrete speech code generation and text sequence modeling respectively. L_{STP} is the loss introduced in the STP module. In the second stage, only the FEC module is optimized, while all other parameters remain frozen. The training objective is

$$\mathcal{L} = \mathcal{L}_{\text{mel}} + \mathcal{L}_{\text{text}}. \tag{12}$$

The two-stage optimization strategy separates speaker representation from emotion modeling, enabling CoCiTTS to achieve stable timbre consistency while preserving emotional expression.

4 CineSpeech Benchmark

We construct CineSpeech, a benchmark specifically designed for character-level speaker consistency evaluation in cross-lingual cinematic dubbing.

Dataset Construction: We collect utterances from 50 publicly available movies and TV series, containing 136 characters and 950 utterances, such as Zhang Fei from *Romance of the Three Kingdoms*, Li Bai from *Chang An*, and Zhu Yingtai from *Butterfly Lovers*. For each character, we collect utterances covering diverse emotions, including angry, happy, sad, surprised, fearful, and disgusted. Such emotion-rich speech enables a comprehensive evaluation of speaker consistency. To ensure data quality, we adopt the automatic preprocessing pipeline of FunCineForge (Liu et al., 2026), followed by manual inspection to remove samples with overlapping speakers, poor recording quality, or ambiguous speaker identities. CineSpeech covers four cross-lingual cinematic dubbing tasks: Chinese-to-English (zh-en), Chinese-to-Korean (zh-ko), Chinese-to-French (zh-fr), and Chinese-to-German (zh-de). For each utterance, the original Chinese transcription is generated by Paraformer (Gao et al., 2022), manually proofread, and translated into the target language. It is worth noting that CineSpeech is only used for evaluation and will be released soon. Details are provided in Appendix A.3.

Consistency Evaluation Metrics: To evaluate character-level speaker timbre consistency, we propose Speaker Consistency (SC). For each character p, we first compute the intra-character $\mathrm{SC}_p^{\text{gen}}$ by averaging the cosine similarity among all generated utterance embeddings from the same character:

$$\mathrm{SC}_p^{\text{gen}} = \frac{2}{N_p(N_p-1)} \sum_{1\le i<j\le N_p} \mathrm{sim}(\mathbf{v}_{p,i}^{\text{gen}}, \mathbf{v}_{p,j}^{\text{gen}}). \tag{13}$$

The overall SC score SC_{gen} is obtained by averaging the consistency scores of all characters:

$$\text{SC}_{\text{gen}} = \frac{1}{P} \sum_{p=1}^{P} \text{SC}_p^{\text{gen}}, \tag{14}$$

where N_p denotes the number of generated utterances for character p, P denotes the total number of characters, and $\text{sim}(\cdot, \cdot)$ represents cosine similarity between speaker embeddings v. We further introduce Speaker Consistency Retention (SCR) to measure the preservation of SC during dubbing:

$$\text{SCR}(\%) = \frac{\text{SC}_{\text{gen}}}{\text{SC}_{\text{ref}}} \times 100\%, \tag{15}$$

where SC_{ref} denotes the SC among all original reference speeches. An SCR above 100% is possible when the generated utterances exhibit a more compact speaker space than the original references.

5 Experiments

5.1 Experimental Settings

Multilingual Extension: We extend IndexTTS2 by further training it with 0.2K hours of Korean, 1.4K hours of French, and 1.6K hours of German speech data from Emilia, obtaining IndexTTS2-ML that supports multilingual TTS: Chinese, English, Korean, French, and German.

Training Setup: We train our model on two NVIDIA RTX PRO 6000 96GB GPUs using the AdamW optimizer for 10 epochs with a batch size of 48, a learning rate of 1×10^{-6}, weight decay of 0.01, and set $\lambda = 0.5$ according to the sensitivity analysis.

Evaluation Datasets: We evaluate four cross-lingual tasks on CineSpeech and an open-source Chinese cinematic dataset CineDub-CN (Liu et al., 2026): zh-en, zh-ko, zh-fr, and zh-de.

Evaluation Metrics: Objectively, WER computed by Whisper (Radford et al., 2023) and UTMOS (Saeki et al., 2022) measure intelligibility and naturalness, respectively. Speaker Similarity (SS) and Speaker Consistency (SC) are computed using ECAPA-TDNN (Desplanques et al., 2020), where SS measures generated-reference similarity, and SC measures character-level consistency among generated utterances. Unless otherwise specified, SC denotes SC_{gen}. Speaker Consistency Retention (SCR) measures the preservation of SC. Emotion Similarity (ES) is calculated using emotion2vec (Ma et al., 2024). Subjectively, Mean Opinion Score (MOS) evaluates overall speech quality and Consistency MOS (CMOS) measures the consistency of speaker timbre among generated utterances. Both are collected from 30 listeners on a five-point scale using 50 samples per TTS system.

Baselines and Training: We compare CoCiTTS with XTTS-v2 (Casanova et al., 2024), CosyVoice3 (Du et al., 2025), and Qwen3-TTS (Hu et al., 2026). For zh-en, CoCiTTS is built upon IndexTTS2 (Zhou et al., 2026) using the Emotional Speech Database (ESD) (Zhou et al., 2021); neither model undergoes multilingual extension. For zh-ko, zh-fr, and zh-de, CoCiTTS is built upon IndexTTS2-ML, and we use the same multilingual checkpoint as the corresponding baseline.

5.2 Experimental Validation

Attribute Leakage: As discussed in **Section 1**, speaker representations extracted from emotionally expressive cinematic speech inevitably contain emotion-related information, which causes speaker timbre inconsistency during cross-lingual dubbing. To verify this observation, we probe whether emotion categories can be predicted from speaker representations extracted by a widely used TTS speaker conditioner (Zhou et al., 2026). Specifically, we extract speaker representations from the Emotional Speech Database (ESD) (Zhou et al., 2021) using a frozen speaker conditioner and train a 5-class emotion classifier (Ravanelli et al., 2021). As shown in Table 1, the original speaker representations achieve 87.3% emotion classification accuracy, indicating substantial emotion leakage (detailed classification results are provided in the Appendix A.2.1). After applying STP, the accuracy decreases to 32.5%, while speaker classification accuracy remains comparable, demonstrating that STP suppresses emotion-related information while preserving speaker identity.

Table 1: Emotion and speaker classification accuracy on speaker representations before and after applying STP, validating the existence of emotion leakage.

Speaker Conditioner	Emotion Acc↓	Speaker Acc↑
Original Conditioner	87.3%	97.9%
+STP	**32.5%**	**98.6%**

Table 2: Experimental validation of Emotion-Induced Timbre Drift on the backbone model.

Dataset	SC_{ref}	Pairs Type	SS	SC
CineSpeech	0.59	Same Emotion	0.51	0.55
		Different Emotion	0.39	0.43
CineDub-CN	0.80	Same Emotion	0.65	0.67
		Different Emotion	0.59	0.62

Furthermore, we investigate whether the FEC module introduces speaker identity leakage. We evaluate FEC-generated emotion representations using speaker and emotion classifiers. These emotion representations achieve 93.8% emotion classification accuracy but only 16.9% speaker classification accuracy, indicating that FEC primarily captures emotion information with limited speaker identity leakage. These results demonstrate that STP and FEC achieve complementary attribute disentanglement, enabling consistent speaker timbre preservation and emotional expression.

Emotion-Induced Timbre Drift: We analyze whether emotional variations cause speaker timbre drift. We conduct validation experiments on both CineSpeech and CineDub-CN (Liu et al., 2026), where 20 speakers are randomly selected from each dataset. For each speaker, we construct same-emotion and different-emotion reference pairs and perform speech synthesis using the backbone model. We then evaluate the generated speech using SS and SC. As shown in Table 2, compared with same-emotion reference pairs, different-emotion pairs result in larger decreases in both SS and SC, demonstrating that emotional variations induce speaker timbre drift. This effect is more evident in CineSpeech, which contains richer and more diverse emotional expressions, where both SS and SC decrease by more than 0.1 under different-emotion pairs.

Comparison between CineSpeech and CineDub-CN: We further compare the SC_{ref} of CineSpeech and CineDub-CN. As shown in Table 2, CineSpeech and CineDub-CN achieve SC_{ref} values of 0.59 and 0.80, respectively, which are used for calculating SCR. This gap in SC_{ref} is mainly due to the richer emotional variations in CineSpeech and affects the absolute range of generated consistency scores in subsequent experiments. In addition, CineSpeech has lower SS and SC in same-emotion pairs compared with CineDub-CN (0.51 vs. 0.65 for SS, and 0.55 vs. 0.67 for SC). This is because, during the construction of CineSpeech, we intentionally select utterances with larger variations even within the same emotion category, including larger variations in emotion intensity, to better reflect the complexity of cinematic speech. Therefore, CineSpeech provides a more challenging benchmark for evaluating the robustness of speaker timbre consistency preservation under emotion-rich cinematic conditions.

5.3 Experiment Results

Performance on CineSpeech: Table 3 shows that CoCiTTS achieves the best SC and SCR across all language tasks. Compared with IndexTTS2 or IndexTTS2-ML, SC increases from 0.42 to 0.47 for zh-en, 0.46 to 0.56 for zh-ko, 0.42 to 0.51 for zh-fr, and 0.55 to 0.63 for zh-de. The improvements demonstrate that suppressing emotion-related variations in speaker conditions effectively alleviates character-level timbre drift under diverse cinematic expressions. SS also improves consistently, indicating that stronger character-level consistency does not sacrifice similarity to the reference speaker. Although ES shows a slight decrease compared with the backbone model, it remains at a competitive and moderate level among all baselines, suggesting that the FEC module successfully preserves emotional expressiveness after speaker purification. CoCiTTS achieves the highest MOS and CMOS, indicating that subjective evaluations are consistent with the trends observed in objective metrics. CoCiTTS also achieves competitive WER and UTMOS scores across all language pairs.

Performance on CineDub-CN: Table 3 shows a similar trend on CineDub-CN. CoCiTTS achieves the highest SS, SC, SCR, and subjective scores. Its WER and UTMOS also improve over the backbone model in most settings, although they are not the best among all systems. Notably, the consistent improvements on both CineSpeech and CineDub-CN demonstrate the robustness of CoCiTTS across different cinematic speech benchmarks.

Table 3: Evaluation results on **CineSpeech** and **CineDub-CN** for cross-lingual zero-shot TTS. For zh-en, Ours and its backbone baseline use IndexTTS2; for other tasks, both use IndexTTS2-ML.

Dataset	Mode	Model	WER(%)↓	UTMOS↑	SS↑	SC↑	SCR(%)↑	ES↑	MOS↑	CMOS↑
CineSpeech	zh-en	XTTS-v2 (Casanova et al., 2024)	5.91	2.47	0.36	0.43	72.88	0.89	3.42	3.21
		CosyVoice3 (Du et al., 2025)	2.05	3.56	0.48	0.39	66.10	0.86	3.71	2.81
		Qwen3-TTS (Hu et al., 2026)	7.54	3.32	0.41	0.34	57.63	0.88	3.52	2.66
		IndexTTS2 (Zhou et al., 2026)	3.15	2.55	0.45	0.42	71.19	**0.90**	3.48	3.06
		Ours	**1.90**	**3.79**	**0.50**	**0.47**	**79.66**	0.88	**4.02**	**3.56**
	zh-ko	XTTS-v2 (Casanova et al., 2024)	54.43	1.72	0.47	0.47	79.66	0.82	2.41	3.53
		CosyVoice3 (Du et al., 2025)	26.71	**2.96**	0.48	0.44	74.58	0.86	3.40	3.34
		Qwen3-TTS (Hu et al., 2026)	14.01	2.64	0.48	0.41	69.49	0.89	3.18	2.98
		IndexTTS2-ML	16.26	1.84	0.47	0.46	77.97	**0.90**	2.72	3.43
		Ours	**12.99**	2.65	**0.49**	**0.56**	**94.92**	0.86	**3.61**	**3.98**
	zh-fr	XTTS-v2 (Casanova et al., 2024)	10.50	1.98	0.39	0.41	69.49	0.86	2.76	3.09
		CosyVoice3 (Du et al., 2025)	6.20	**2.84**	0.42	0.39	66.10	0.90	3.35	2.70
		Qwen3-TTS (Hu et al., 2026)	12.67	2.59	0.44	0.37	62.71	**0.92**	3.09	2.65
		IndexTTS2-ML	7.98	1.69	0.41	0.42	71.19	0.90	2.61	2.96
		Ours	**5.66**	2.41	**0.47**	**0.51**	**86.44**	0.88	**3.47**	**3.87**
	zh-de	XTTS-v2 (Casanova et al., 2024)	22.29	1.92	0.40	0.45	76.27	0.85	2.78	3.72
		CosyVoice3 (Du et al., 2025)	2.60	**2.91**	0.43	0.42	71.19	0.87	3.52	3.15
		Qwen3-TTS (Hu et al., 2026)	4.80	2.86	0.40	0.39	66.10	0.80	3.31	2.74
		IndexTTS2-ML	4.57	1.77	0.37	0.55	93.22	**0.88**	2.75	3.88
		Ours	**2.15**	2.49	**0.44**	**0.63**	**106.78**	**0.88**	**3.56**	**4.15**
CineDub-CN	zh-en	XTTS-v2 (Casanova et al., 2024)	**1.59**	3.28	0.52	0.44	55.00	0.82	3.36	3.75
		CosyVoice3 (Du et al., 2025)	2.53	4.15	0.64	0.56	70.00	0.86	3.82	3.95
		Qwen3-TTS (Hu et al., 2026)	4.69	**4.23**	0.58	0.56	70.00	0.84	3.76	3.91
		IndexTTS2 (Zhou et al., 2026)	2.69	3.63	0.62	0.62	77.51	**0.88**	3.71	4.02
		Ours	2.67	3.71	**0.65**	**0.70**	**87.50**	0.84	**4.05**	**4.20**
	zh-ko	XTTS-v2 (Casanova et al., 2024)	26.06	2.23	0.69	0.58	72.50	0.84	2.85	3.98
		CosyVoice3 (Du et al., 2025)	**4.17**	3.30	0.72	0.72	90.00	0.86	3.55	4.32
		Qwen3-TTS (Hu et al., 2026)	4.65	**3.48**	0.72	0.65	81.25	0.88	3.61	4.16
		IndexTTS2-ML	8.59	1.96	0.68	0.65	81.25	**0.90**	2.82	4.11
		Ours	7.14	2.44	**0.77**	**0.76**	**95.00**	**0.90**	**3.72**	**4.46**
	zh-fr	XTTS-v2 (Casanova et al., 2024)	2.75	2.40	0.50	0.48	60.00	0.84	3.01	3.62
		CosyVoice3 (Du et al., 2025)	1.71	3.28	0.60	0.71	88.75	0.86	3.63	4.22
		Qwen3-TTS (Hu et al., 2026)	6.30	**3.50**	0.62	0.58	72.50	**0.88**	3.58	3.65
		IndexTTS2-ML	3.48	2.36	0.59	0.62	77.50	0.86	3.19	3.96
		Ours	**1.42**	2.58	**0.63**	**0.74**	**92.50**	0.86	**3.79**	**4.37**
	zh-de	XTTS-v2 (Casanova et al., 2024)	3.69	2.43	0.47	0.70	87.50	0.83	3.18	4.19
		CosyVoice3 (Du et al., 2025)	0.88	3.48	0.57	0.72	90.00	0.86	3.66	4.28
		Qwen3-TTS (Hu et al., 2026)	9.81	**3.69**	0.53	0.62	77.50	0.86	3.42	4.04
		IndexTTS2-ML	1.69	2.67	0.55	0.65	81.25	**0.88**	3.38	4.12
		Ours	**0.76**	2.76	**0.59**	**0.76**	**95.00**	0.87	**3.83**	**4.39**

5.4 Ablation Study

We conduct ablation studies on CineSpeech and CineDub-CN to validate the effectiveness of the proposed STP and FEC modules. We report SC and ES, averaged over four cross-lingual cinematic dubbing tasks. As shown in Figure 4 (a) and (b), STP raises SC from 0.46 to 0.52 on CineSpeech and from 0.64 to 0.71 on CineDub-CN, demonstrating its effectiveness in reducing emotion-related variations in speaker conditions and enhancing character-level speaker consistency. Meanwhile, STP leads to a slight decrease in ES, which is expected since emotion-related information is partially suppressed by the STP module. In contrast, FEC compensates for the emotional information removed by STP and restores the emotional expressiveness of generated speech. Specifically, +STP+FEC achieves ES values close to the backbone model while further improving SC. These results demonstrate that CoCiTTS effectively balances character-level speaker consistency and emotional expressiveness through the collaboration of STP and FEC.

5.5 Sensitivity Analysis of Hyperparameter λ

We investigate the effect of $\lambda \in \{0, 0.25, 0.5, 0.75, 1.0\}$ on CineSpeech and CineDub-CN. As shown in Figure 5, increasing λ improves SC by suppressing emotion-related variations, while overly large λ reduces SS because excessive removal of emotion-related information may affect speaker identity representations. Therefore, we select $\lambda = 0.5$ as the optimal trade-off between SS and SC.

5.6 Generalization

To evaluate the generalizability of the proposed STP module, we integrate it into another zero-shot TTS backbone, CosyVoice (Du et al., 2024a). We choose CosyVoice because it adopts reference

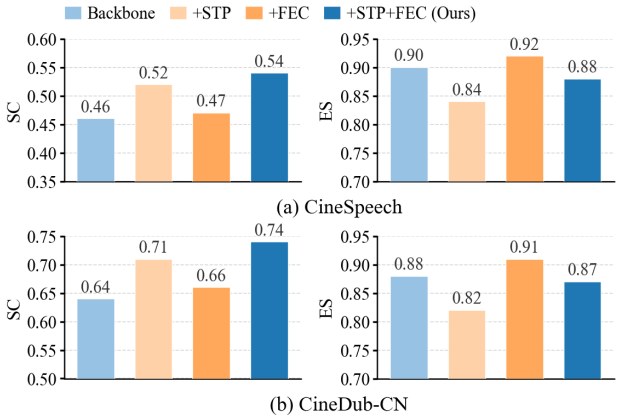


Figure 4: Ablation studies on CineSpeech and CineDub-CN.

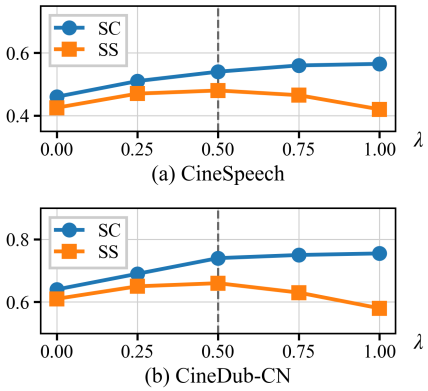


Figure 5: Sensitivity analysis of hyper-parameter λ.

speech as input to its Text-to-Semantic module and supports cross-lingual TTS. Since CosyVoice does not support French and German synthesis, we only evaluate the zh-en and zh-ko tasks.

As shown in Table 4, STP improves SC on both datasets. On CineSpeech, SC increases from 0.36 to 0.42 for zh-en and from 0.40 to 0.45 for zh-ko; the corresponding SCR gains are 10.17 and 8.47 percentage points. On CineDub-CN, SC increases from 0.64 to 0.69 for zh-en and from 0.70 to 0.75 for zh-ko, with SCR gains of 6.25 percentage points in both tasks. STP also improves SS, MOS, and CMOS in all four settings while reducing WER. UTMOS improves in three settings and changes only slightly on CineDub-CN zh-en, from 3.77 to 3.75.

Overall, these results demonstrate that STP is not limited to the CoCiTTS architecture and can effectively improve speaker consistency when applied to different zero-shot TTS backbones. This validates the general applicability of STP for learning more robust and consistency-oriented speaker representations for cross-lingual cinematic dubbing.

Table 4: Generalization of the STP module on CosyVoice across CineSpeech and CineDub-CN.

Dataset	Mode	Model	WER(%)↓	UTMOS↑	SS↑	SC↑	SCR(%)↑	MOS↑	CMOS↑
CineSpeech	zh-en	CosyVoice	5.72	3.00	0.42	0.36	61.02	3.41	2.75
		+STP	**3.17**	**3.33**	**0.45**	**0.42**	**71.19**	**3.56**	**2.93**
	zh-ko	CosyVoice	14.32	2.41	0.49	0.40	67.80	3.05	3.06
		+STP	**13.26**	**2.69**	**0.53**	**0.45**	**76.27**	**3.27**	**3.40**
CineDub-CN	zh-en	CosyVoice	3.01	**3.77**	0.50	0.64	80.00	3.47	4.15
		+STP	**2.85**	3.75	**0.56**	**0.69**	**86.25**	**3.50**	**4.25**
	zh-ko	CosyVoice	6.87	2.81	0.63	0.70	87.50	3.50	4.33
		+STP	**5.40**	**2.85**	**0.69**	**0.75**	**93.75**	**3.69**	**4.42**

6 Conclusion

In this paper, we observe the character-level speaker inconsistency problem in cross-lingual cinematic dubbing and identify that emotion leakage in speaker representations is an important factor. To address this issue, we propose CoCiTTS, a zero-shot TTS framework that combines Speaker Timbre Purification for consistent speaker conditioning with Fine-grained Emotion Compensation to restore emotional cues through an independent pathway. To better evaluate speaker consistency in cinematic scenarios, we construct CineSpeech, an emotionally diverse cinematic speech benchmark, and introduce speaker consistency metrics. Experiments on multiple cross-lingual tasks demonstrate that CoCiTTS effectively improves character-level speaker consistency while maintaining competitive speech quality and emotional expression. Overall, this study provides a new perspective for addressing speaker timbre drift in zero-shot TTS and advances the development of consistent and expressive cross-lingual cinematic dubbing systems.

AI USE STATEMENT

In this work, we used generative AI tools to improve the readability and grammar of this paper. We also used AI tools to process parts of the dataset, including automated translation and text preprocessing. These tools were not used to develop the research methodology, implement the proposed method, analyze the experimental results, or interpret the findings. We reviewed the AI-assisted text and dataset outputs and verified their accuracy and suitability for this study. We take full responsibility for the final content of this work, including any text, claims, data, or artifacts produced with the assistance of AI tools.

REPRODUCIBILITY STATEMENT

We support reproducibility by detailing the CoCiTTS architecture in Section 3, including the Speaker Timbre Purification (STP) and Fine-grained Emotion Compensation (FEC) modules and the two-stage training strategy; the overall framework is illustrated in Figure 2. Section 5.1 describes the experimental settings, including the multilingual extension, datasets, evaluation metrics, baselines, and training setup. Section 4 presents the CineSpeech benchmark and speaker-consistency metrics. Appendix A.2 provides supplementary experiments, and Appendix A.3 details the CineSpeech construction and processing pipeline. Upon publication, we will publicly release the code and CineSpeech dataset to support replication and further research.

REFERENCES

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. Xtts: a massively multilingual zero-shot text-to-speech model. In *Interspeech 2024*, pp. 4978–4982, 2024. doi: 10.21437/Interspeech.2024-2016.

Qi Chen, Mingkui Tan, Yuankai Qi, Jiaqiu Zhou, Yuanqing Li, and Qi Wu. V2c: Visual voice cloning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21210–21219, 2022. doi: 10.1109/CVPR52688.2022.02056.

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. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718, 2025. doi: 10.1109/TASLPRO.2025.3530270.

Marta Costa-jussa, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Gonzalez, Prangthip Hansanti, John Hoffman, and Jeff Wang. Scaling neural machine translation to 200 languages. *Nature*, 630, 06 2024. doi: 10.1038/s41586-024-07335-x.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2019. doi: 10.1109/CVPR.2019.00482.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. 10 2020. doi: 10.21437/Interspeech.2020-2650.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens, 2024a. URL https://arxiv.org/abs/2407.05407.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jing-Ru Zhou. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *ArXiv*, abs/2412.10117, 2024b. URL https://api.semanticscholar.org/CorpusID:274762932.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, Shiliang Zhang, Wen Wang, and Jieping Ye. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training, 2025. URL https://arxiv.org/abs/2505.17589.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030, January 2016. ISSN 1532-4435.

Zhifu Gao, Shiliang Zhang, Ian Mcloughlin, and Zhijie Yan. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. pp. 2063–2067, 09 2022. doi: 10.21437/Interspeech.2022-9996.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-tts technical report, 2026. URL https://arxiv.org/abs/2601.15621.

Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: Adaptive curriculum learning loss for deep face recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5900–5909, 2020. doi: 10.1109/CVPR42600.2020.00594.

Ziyue Jiang, Jinglin Liu, Yi Ren, Jinzheng He, Zhenhui Ye, Shengpeng Ji, Qian Yang, Chen Zhang, Pengfei Wei, Chunfeng Wang, Xiang Yin, Zejun MA, and Zhou Zhao. Mega-TTS 2: Boosting prompting mechanisms for zero-shot speech synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mvMI3N4AvD.

Ziyue Jiang, Yi Ren, Ruiqi Li, Shengpeng Ji, Boyang Zhang, Zhenhui Ye, Chen Zhang, Bai Jionghao, Xiaoda Yang, Jialong Zuo, Yu Zhang, Rui Liu, Xiang Yin, and Zhou Zhao. Megatts 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis, 2025. URL https://arxiv.org/abs/2502.18924.

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. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *ICML*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22605–22623. PMLR / OpenReview.net, 2024. URL http://dblp.uni-trier.de/db/conf/icml/icml2024.html#JuWS0XYLLST000024.

Jiaxuan Liu, Yang Xiang, Han Zhao, Xiangang Li, and Zhenhua Ling. Funcineforge: A unified dataset toolkit and model for zero-shot movie dubbing in diverse cinematic scenes, 2026. URL https://arxiv.org/abs/2601.14777.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15747–15760, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.931. URL https://aclanthology.org/2024.findings-acl.931/.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris,

Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. Speechbrain: A general-purpose speech toolkit, 2021. URL https://arxiv.org/abs/2106.04624.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. In *Interspeech 2022*, pp. 4521–4525, 2022. doi: 10.21437/Interspeech.2022-439.

Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yichong Leng, Lei He, Tao Qin, sheng zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Rc7dAwVL3v.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018. doi: 10.1109/ICASSP.2018.8461375.

Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pp. 3927–3935, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386517. doi: 10.1145/3474085.3475587. URL https://doi.org/10.1145/3474085.3475587.

Jenthe Thienpondt and Kris Demuynck. Ecapa2: A hybrid neural network architecture and training strategy for robust speaker embeddings. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2023. doi: 10.1109/ASRU57964.2023.10389750.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. Cam++: A fast and efficient network for speaker verification using context-aware masking. pp. 5301–5305, 08 2023a. doi: 10.21437/Interspeech.2023-1513.

Ju-Chiang Wang, Wei-Tsung Lu, and Minz Won. Mel-band roformer for music source separation, 2023b. URL https://arxiv.org/abs/2310.01809.

Zhedong Zhang, Liang Li, Gaoxiang Cong, Chunshan Liu, Yuhan Gao, Xiaowan Wang, Tao Gu, and Yuankai Qi. Instructdubber: instruction-based alignment for zero-shot movie dubbing. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i15.38298. URL https://doi.org/10.1609/aaai.v40i15.38298.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 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)*, pp. 920–924. IEEE, 2021.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. Indextts2: a breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i41.40820. URL https://doi.org/10.1609/aaai.v40i41.40820.

A Appendix

A.1 Related Work

Speaker Representation Learning. Speaker representation learning focuses on extracting speaker identity while suppressing irrelevant variations. Methods such as x-vector (Snyder et al., 2018)

and ECAPA-TDNN (Desplanques et al., 2020) improve the speaker discriminability and robustness. Gradient Reversal Layer (GRL) (Ganin et al., 2016) based adversarial learning has been widely used to suppress attribute-related variations.

A.2 Supplementary Experiments

A.2.1 Emotion Leakage

To provide a more detailed analysis of emotion leakage in speaker representations, we report the confusion matrix of the five-class emotion classifier trained on the original speaker representations extracted by the frozen speaker conditioner. As shown in Figure 6, the classifier achieves a mean accuracy of 87.3%, with per-class recall ranging from 85% to 89%. The high diagonal values show that emotion can be reliably predicted from speaker representations, confirming substantial emotion leakage that contributes to speaker timbre drift.



Figure 6: Confusion matrix of the five-class emotion classifier trained on the original speaker representations. The average classification accuracy (87.3%) demonstrates substantial emotion leakage in the original speaker conditioning.

A.2.2 Emotion-Controlled Intervention

To further verify whether emotion leakage in speaker representations contributes to timbre inconsistency, we conduct an emotion-controlled intervention experiment. We construct reference pairs from the ESD (Zhou et al., 2021) dataset, where each pair consists of utterances from the same speaker with identical linguistic content but different emotions. During synthesis, we fix the target text and the reference speech supplied to the emotion conditioner, while only the emotional state of the reference speech supplied to the speaker conditioner is varied. We compare IndexTTS2-ML and CoCiTTS by measuring the similarity of extracted speaker conditions (Cond-Sim) and generated speeches (Speaker Consistency, SC). As illustrated in Table 5, CoCiTTS achieves higher Cond-Sim and SC than IndexTTS2-ML, indicating that the Speaker Timbre Purification (STP) module effectively reduces emotion-induced variations in speaker representations and improves speaker timbre consistency.

Model	**Cond-Sim↑**	**SC↑**
IndexTTS2-ML	0.71	0.68
CoCiTTS	**0.86**	**0.79**

Table 5: Emotion-controlled intervention results for evaluating the impact of emotion leakage on speaker conditions and speaker timbre consistency.

A.2.3 Inter-Speaker Discriminability Analysis

This subsection aims to evaluate the discriminability between utterances of different speakers. Figure 7 visualizes the t-SNE projections of ECAPA2 (Thienpondt & Demuynck, 2023) embeddings for 195 utterances from 15 speakers, randomly sampled from CineSpeech and CineDub-CN. We use ECAPA2 for visualization to verify that the observed inter-speaker discriminability generalizes to a different and more recent speaker embedding extractor. The same colors denote the same speaker, while the stars indicate randomly selected enrollment utterances. n denotes the number of utterances of each speaker. The utterances of most speakers form relatively tight clusters with distinct margins between them. This demonstrates that our model maintains robust inter-speaker discriminability while enhancing character-level speaker consistency.

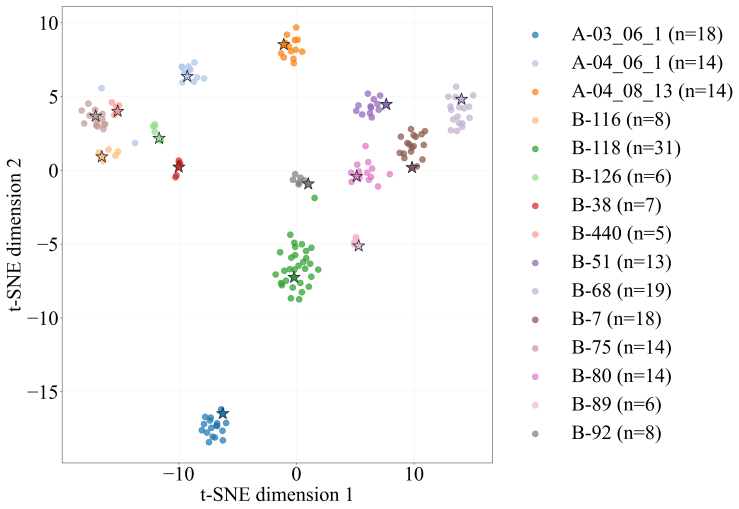


Figure 7: t-SNE visualization of speaker embeddings.

A.3 CineSpeech Dataset Construction

As illustrated in Figure 8, we process film and television clips and construct the CineSpeech dataset through an open-source tool of (Liu et al., 2026), followed by elaborate and careful manual proofreading. Details are as follows:

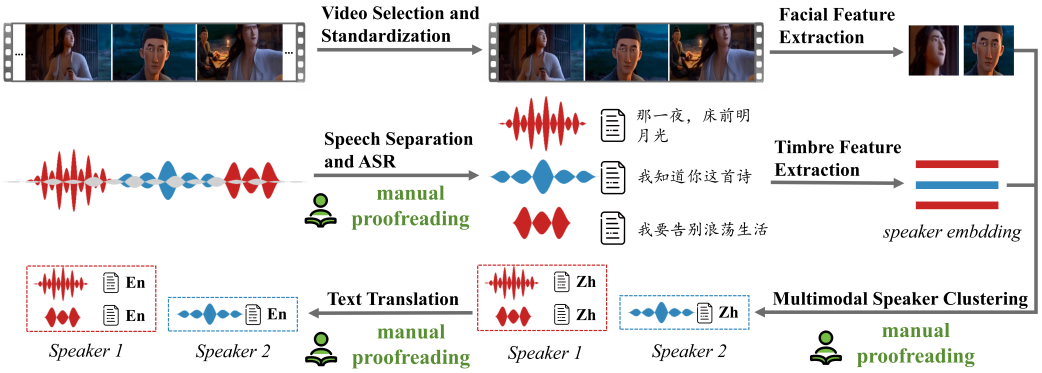


Figure 8: CineSpeech dataset construction pipeline.

Video selection and standardization. We select dialogue-rich clips with clear speech, visible characters, and relatively little background noise. The beginning and end segments containing irrelevant content are removed.

Speech separation and ASR. We use Mel-Band RoFormer (Wang et al., 2023b) to separate speech from background sound. Then we use FSMN to locate speech regions, Paraformer (Gao et al., 2022) to transcribe Chinese speech, and CT-Transformer to restore punctuation. The recordings are then divided into utterance-level clips.

Multimodal speaker clustering. We perform automatic speaker clustering using both audio and visual cues. On the audio side, CAM++ (Wang et al., 2023a) is used to extract speaker embeddings. On the visual side, TalkNet (Tao et al., 2021) identifies the active speaker. CurricularFace (Huang et al., 2020) provides facial identity features, and FAN facial landmarks provide information about facial and lip regions. Speaker embeddings and visual identity features are combined to assign a speaker label to each utterance. The resulting samples are then organized by speaker identity. Each record contains an audio clip, its transcript, a language label, and a speaker identifier.

Text translation. The verified Chinese transcripts are translated into the target languages by NLLB-200 (Costa-jussa et al., 2024) to form the cross-lingual evaluation tasks. Figure 8 illustrates the zh-en task as an example.

Manual quality control. We manually review the automatically processed samples and remove clips with incorrect speaker assignments, transcription errors, excessive noise, overlapping speakers, or incomplete utterance boundaries.