Neurocomputing 670 (2026) 132529

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Controllable timbre cloning and style replication with reference speech examples for multimodal human-computer interaction

Tianwei Lan [a], Yuhang Guo [a], Mengyuan Deng [b], Jing Wang [b,*], Wenwu Wang [c], Chong Feng [a]

[a] School of Computer Science and Technology, Beijing Institute of Technology, No. 5, Zhongguancun South Street, Haidian District, Beijing, 100081, China
[b] School of Information and Electronics, Beijing Institute of Technology, No. 5, Zhongguancun South Street, Haidian District, Beijing, 100081, China
[c] The Centre for Vision, Speech, and Signal Processing, University of Surrey, GU2 7XH, Guildford, UK

HIGHLIGHTS

- Proposed a new speech task using dual-reference audio for precise timbre/style control & novel combinations.
- Control-TTS model recombines speech traits for new timbre-style mixes, matching SOTA performance.
- Multi-encoder design validated: effectively recombines timbre/style features via ablation studies & t-SNE.

ARTICLE INFO

Communicated by R. Yang

Keywords:
Multimodal human-computer interaction
Timbre cloning and style replication
Controllable speech synthesis

ABSTRACT

Natural and personalized speech interaction is one of the core requirements for advancing Multimodal Human-Computer Interaction (HCI), with applications widely seen in smart home devices, voice assistants, and mobile devices. In recent years, the demand for speech in the HCI field has shifted from basic speech generation to precise customization of speaker timbre and speaking style, aiming to achieve more intuitive and immersive multimodal human-computer interaction. However, existing speech personalization technologies have significant limitations: zero-shot speech synthesis methods lack the capability for style control, while traditional style-controllable synthesis methods fail to accurately specify speaker timbre, making it difficult to balance personalization between speaker timbre and speaking style. To address this issue, we define a new task: Controllable Timbre Cloning and Style Replication with Reference Speech Examples. This task aims to directly control speaker timbre and speaking style through two reference speech examples, allowing timbre cloning and style replication to generate new timbre-style combinations. To tackle this task, we propose the Control-TTS model. This model utilizes distinct reference speeches to separately control the timbre and speaking style features of the speaker in the synthesized audio, enabling free combinations of timbre and style. This approach generates synthetic speech with rich expressivity, providing a more flexible and customizable solution for speech personalization in HCI scenarios. Our experiments on the VccmDataset demonstrate that Control-TTS achieves comparable or state-of-the-art performance in terms of metrics such as naturalness mean opinion score (NMOS), word error rate (WER), speaker similarity, and style similarity. Our demo is available at https://progressivetts.github.io/Control_TTS/.

1. Introduction

Natural and personalized speech interaction serves as the cornerstone for advancing multimodal human-computer interaction (HCI) [1–3], with applications that permeate smart home devices, voice assistants, and mobile devices. This technology bridges the modal gap between text and audio, converting text-based information into expressive and rich speech output. It is an essential interface for intuitive multimodal human-computer interaction. As speech interaction technologies evolve, the naturalness, expressiveness, and emotional adaptability of generated speech have improved significantly, enabling compelling performances in scenarios such as storytelling, virtual assistance, and educational applications.

* Corresponding author.
Email addresses: 3220231230@bit.edu.cn (T. Lan), guoyuhang@bit.edu.cn (Y. Guo), 3120230720@bit.edu.cn (M. Deng), wangjing@bit.edu.cn (J. Wang), w.wang@surrey.ac.uk (W. Wang), fengchong@bit.edu.cn (C. Feng).

https://doi.org/10.1016/j.neucom.2025.132529
Received 1 August 2025; Received in revised form 18 November 2025; Accepted 24 December 2025
Available online 30 December 2025

In recent years, HCI demands have shifted beyond basic fluent speech generation toward the precise customization of speaker timbre and speaking style. Users now seek not only intelligible speech output but also the ability to tailor vocal characteristics to specific personas and contextual styles. This shift presents new challenges for personalized speech technology. The ideal solution would allow independent control over speaker timbre (capturing the unique vocal identity of a speaker) and speaking style (reflecting expressive nuances such as emotion or intonation) while preserving the accuracy of the content.

Currently, several research approaches have emerged in controllable speech synthesis, but each has its own limitations. Zero-shot speech synthesis [4–7] can only transfer the emotion-timbre combination from reference speech, without the ability to freely specify arbitrary styles or modify the emotional expression of reference speech during synthesis [8–11]. Although several studies employ emotion IDs to control emotion types [12], this approach only enables coarse-grained emotional control and does not capture subtle variations within the same emotion category. Similarly, style-controllable speech synthesis systems [13–15] can only modulate speech styles while being unable to specify speaker timbre. Although speaker IDs have been adopted for timbre control in several works [16], this approach is constrained by limited speaker diversity and fails to generate sufficiently varied timbres. The fundamental flaw of these two types of work lies in their inability to effectively decouple timbre and style, resulting in insufficient control over synthesized speech.

In addition, some works use text descriptions to control the speaking style of synthesized speech [13,14], or first extract text descriptions from style reference audio and then use the extracted text descriptions to guide the synthesis of stylized speech [17,18]. However, describing speech styles with words is insufficient to fully and comprehensively represent the style patterns. The lack of accuracy in descriptions can lead to the failure to accurately convey information about the original speech style, resulting in differences between the generated speech style and that of the original audio. Meanwhile, this process also increases the operational difficulty of the system, imposes a burden on users, and hinders the promotion and popularization of this technology. These limitations stem from a core oversight: treating style as a text-encodable attribute rather than a distinct perceptual feature that requires independent modeling.

To address these inherent limitations, our proposed Control-TTS directly controls the speaker's timbre and speaking style of the synthesized speech through reference audio, and can specify arbitrary speaking content, as shown in Fig. 1. Unlike single-reference zero-shot TTS, our dedicated speaker encoder separates speaker-specific features from the reference speech, while the style encoder independently captures prosodic and emotional nuances, thereby eliminating the binding between timbre and style. Compared with text description-based methods, our design does not require explicit text descriptions. By directly modeling style as perceptual embeddings, it preserves fine-grained style details and reduces the burden on users. For methods based on discrete emotion IDs, our continuous style embedding space naturally supports subtle style variations because it learns from perceptual similarity rather than relying on predefined categories.

Our contributions are as follows:

- We identify the demand for personalized customization of speaker timbre and style in the field of multimodal human-computer interaction, establishing the task of Controllable Timbre Cloning and Style Replication with Reference Speech Examples. This task leverages two reference speech samples to control both speaker timbre and speaking style more precisely, enabling the generation of novel timbre-style combinations.
- We propose Control-TTS, a novel model that generates new combinations of speaker timbre and speaking style by effectively recombining characteristics from speech examples. Experimental results demonstrate that Control-TTS exhibits comparable or state-of-the-art performance on this task.
- We conduct comparative experiments on the number of encoders, t-SNE clustering, and other experiments. We also explore the impact of reference audio length, background noise, training data volume, and cross-lingual reference audio on the quality of synthesized audio. These experiments demonstrate that Control-TTS uses multiple encoders to model speech from multiple perspectives, enabling more effective extraction of timbre and style from audio during speech synthesis.

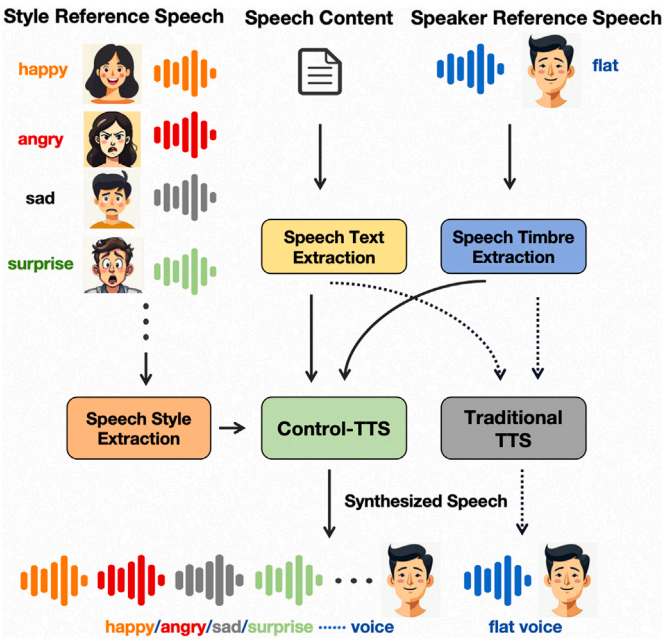


Fig. 1. Schematic diagram of Control-TTS. Control-TTS extracts the timbre of speaker reference speech and the style of style reference speech and synthesizes the final audio. Traditional TTS models can only adopt the style and emotion of the original reference speech.

2. Related work

We propose the task of Controllable Timbre Cloning and Style Replication with Reference Speech Examples. The related work of this new task includes zero-shot speech synthesis, style-controllable speech synthesis, and previous controllable speech synthesis tasks under the guidance of text descriptions. Below, we introduce previous works in these three parts and explain the differences between our work and these existing works.

2.1. Zero-shot TTS

Zero-shot speech synthesis refers to synthesizing the voice of an unseen speaker with only the guidance of a few seconds of voice prompts. This technology is also called voice cloning. With the introduction of different model architectures, the effectiveness of zero-shot speech synthesis is also continuously improving. VALL-E [4] uses a discrete codec representation to combine autoregressive and non-autoregressive models in a cascade manner, retaining the powerful contextual functionality of the language model. NaturalSpeech2 [5] replaces discrete neural codec tags with continuous vectors, introducing in-context learning into the diffusion model and further improving the quality of synthesized speech. Mega-TTS [6] uses traditional Mel-spectrograms to decouple timbre and prosody and employs autoregressive methods to further model prosody. VoiceBox [7] is a non-autoregressive stream matching model that is trained to fill in speech given speech context and text. It is worth noting that although zero-shot speech synthesis has made great

Table 1
Functional comparison of Control-TTS and other TTS systems.

Models	Timbre Clone	Style Control	Integration of Timbre and Style
VALL-E[4]	✔	✘	✘
Mega-TTS[6]	✔	✘	✘
VoiceBox[7]	✔	✘	✘
StyleTTS2[21]	✔	✘	✘
PromptTTS[13]	✘	✔	✘
PromptStyle[22]	✘	✔	✘
InstructTTS[14]	✘	✔	✘
Textrolspeech[15]	✘	✔	✘
Unistyle[17]	✔	✔	✔ (style controlled by text description)
ControlSpeech[18]	✔	✔	✔ (style controlled by text description)
Control-TTS (Ours)	✔	✔	✔ **(style controlled by audio reference)**

progress, this technology can only control the timbre of the synthesized speech, but not the style of the speech. In contrast, our Control-TTS achieves control over both timbre and style at the same time, thereby addressing this limitation.

2.2. Style-controllable speech synthesis

Several studies control the style of synthesized speech through text prompts. The text description usually includes gender, pitch, speaking speed, emotion, etc. In this type of study, the model understands the text description and converts it into the style of synthesized speech, which has certain cross-modal capabilities. PromptTTS [13] uses manually annotated text prompts to describe the five attributes of speech (gender, pitch, speaking speed, energy, and emotion) and trains the model on two synthetic speaker datasets and LibriTTS [19]. InstructTTS [14] uses a three-stage training method to capture semantic information from natural language style prompts and uses the semantic information as conditional input for the TTS system. Textrolspeech [15] regards style-controllable TTS as a language modeling task and uses a codec architecture based on VALL-E [4]. PromptTTS2 [20] proposes using LLM to automatically create text descriptions of speech style and adopts a diffusion model to capture the one-to-many relationship between speech and text descriptions. It is worth noting that existing style-controllable speech synthesis systems are either fixed-speaker speech synthesis systems or can only specify a limited number of timbres through SpeakerID, lacking the capability of timbre cloning. Our Control-TTS can specify the timbre of any speaker through reference speech.

2.3. Speaker-specific and style-controllable TTS

To address the limitations of zero-shot TTS and style-controllable speech synthesis, several studies suggest using textual descriptions to control the speaking style or the speaker's timbre. Unistyle [17] uses two reference speech samples to control the language style and speaker timbre, respectively. The speaker's timbre is directly controlled by the timbre reference speech, and the style reference speech must be converted into a text description before being re-entered into the model. After that, the style of the synthesis speech is controlled, which is equivalent to a cascade synthesis process. ControlSpeech [18] directly controls the synthesized timbre through timbre reference speech, but this model also controls the style of speech through text description.

We argue that describing speech style in words is not enough to fully and comprehensively reflect the style, during which the speech style may lose accuracy due to inaccurate or incomplete descriptions, which in turn causes the synthesized speech style to deviate from the style of the original speech. At the same time, for users, a relatively complete and comprehensive description of the speech style is required to synthesize an ideal speech. This requirement raises the bar for deploying the system and is not conducive to the widespread dissemination of related technologies. Our Control-TTS proposed in this paper directly controls the speaking style through reference speech, avoiding this error while improving the convenience of using the model.

As shown in Table 1, we present a functional comparison between Control-TTS and other speech synthesis systems. Compared to other models, Control-TTS is capable of simultaneously achieving voice cloning and style control. Moreover, during the style control process, it utilizes audio references, which can more accurately describe style details compared to text descriptions.

3. Proposed methods

In this section, we first introduce the overall structure and inference process of Control-TTS, demonstrating how the model achieves controllable speech synthesis in terms of speaker style and timbre. Then, we describe the role of each module in the inference process, along with its inputs and outputs. Finally, we discuss the training process of the model and the rationale behind this training approach.

3.1. Overview

Fig. 2 shows the overall architecture of Control-TTS. The model's inputs include:

(1) A speaker reference speech R_{spk} used to control the timbre of the synthesized speech.
(2) A style reference speech R_{sty} used to control the speaking style of the synthesized speech.
(3) A phoneme sequence $P = [P_1, P_2, \ldots, P_n]$ derived from text, where n is the length of the phoneme, used to control the content of the synthesized speech.

The model's output is the synthesized speech generated under the control of these three inputs.

Our processing pipeline for raw speech samples comprises six key stages. First, audio samples are read exclusively in mono-channel format. Second, all speech signals are uniformly resampled to 24 kHz. Third, audio length standardization is performed: segments shorter than 0.6 s are zero-padded to meet the minimum duration threshold, while all recordings are prefixed and suffixed with 5000 zero-valued samples (silence segments). Fourth, Mel-spectrograms are extracted using 80 Mel-frequency bands. Fifth, spectral normalization is applied through logarithmic compression and standardization of the Mel-spectrograms. Finally, during training, reference audio segments undergo random cropping with a maximum Mel-spectrogram frame length constrained to 192 frames.

The model's Speaker Encoder encodes R_{spk} to extract the timbre embedding S_{t}. Subsequently, S_{t} will be fed into the Decoder to guide the synthesis of the audio timbre.

The model's Prosody Encoder and Duration Encoder process R_{sty} to generate embeddings for the prosody and duration predictors within

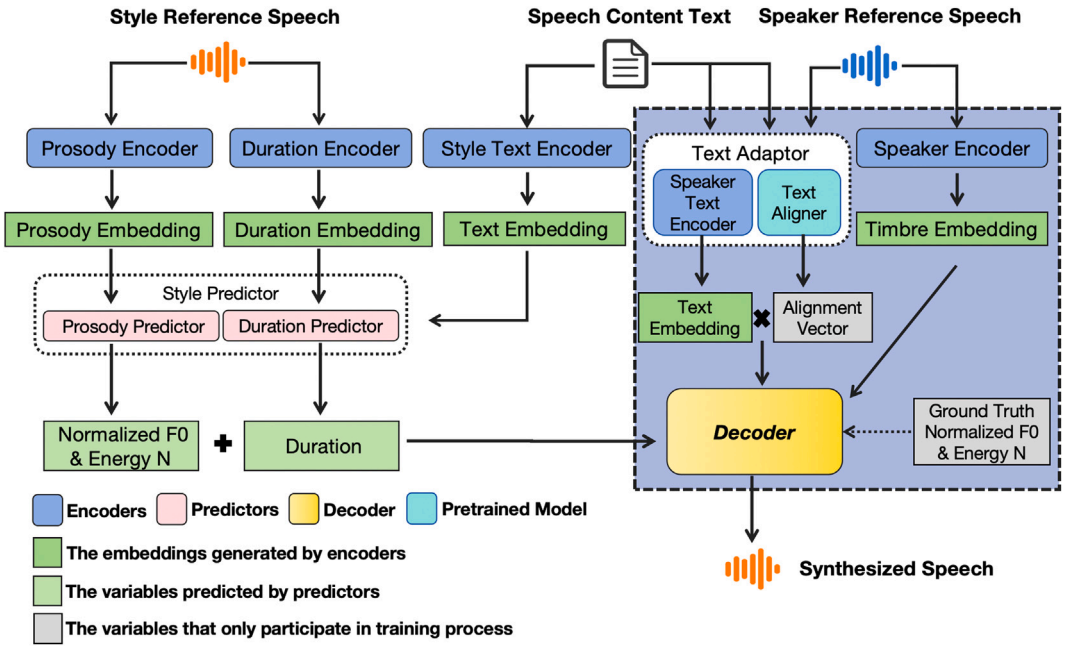


Fig. 2. The overall structure of Control-TTS. In the first stage of training, only the modules in the dashed box with a blue background participate, and in the second stage is the joint training of all modules occurs. The dotted arrows only participate in the first stage of training and do not participate in the second stage.

the style predictor. The style text encoder additionally provides textual encodings to the style predictor, which estimates phoneme durations, normalized fundamental frequency (F0), and energy (E) trajectories. These acoustic features exhibit strong content dependence in speech synthesis. The predicted durations, F0, and energy parameters are subsequently fed to the Decoder for style-controlled speech generation.

The Text Adaptor consists of two components: the Speaker Text Encoder and the Text Aligner. The Speaker Text Encoder converts phoneme sequences into phoneme embeddings. During inference, phoneme durations are predicted by the Duration Predictor, while training employs the Text Aligner to extract ground-truth durations from reference speech. These durations form an alignment matrix A^{n*l}, where n denotes the phoneme sequence length and l represents the Mel-spectrogram frame count. The phoneme embeddings are then aligned through matrix multiplication with A before decoder integration, enabling accurate audio content synthesis.

This describes the Control-TTS framework for controllable speech synthesis. The subsequent sections detail the individual sub-modules.

The above is the overall process of Control-TTS for controllable speech synthesis. In the following sections, we will introduce the details of each sub-module.

3.2. Control-TTS module introduction

3.2.1. Encoder

Control-TTS contains three types of encoders, which respectively encode R_{sty}, P, and R_{spk}. The following sections introduce each of these components in detail.

Style Speech Encoder

The Style Speech Encoder employs parallel prosody and duration encoders to extract prosodic (S_p) and duration (S_d) representations from reference Mel-spectrograms M_{sty}, capturing normalized F0, energy (E), and phoneme duration information. The architecture implements a hierarchical residual network with four bottleneck residual blocks featuring spectral-normalized convolutional layers and instance normalization. Global adaptive pooling generates 512-channel features, projected to 128-dimensional embeddings S_p and S_d. The specific structure of this module is shown in the Fig. 3(a).

Speaker Encoder

The Speaker Encoder's function is to extract the speaker's timbre information. Given the Mel-spectrogram of R_{sty}, it provides a speaker embedding S_t. Its structure is the same as the style encoder, and it also includes four layers of residual networks with a bottleneck structure. The specific structure of this module is shown in the Fig. 3(a).

Style Text Encoder

The Style Text Encoder consists of a pre-trained phoneme-level Bert [23] and a linear layer, with which we can obtain a fine-grained text embedding of the phoneme context.

3.2.2. Text adaptor

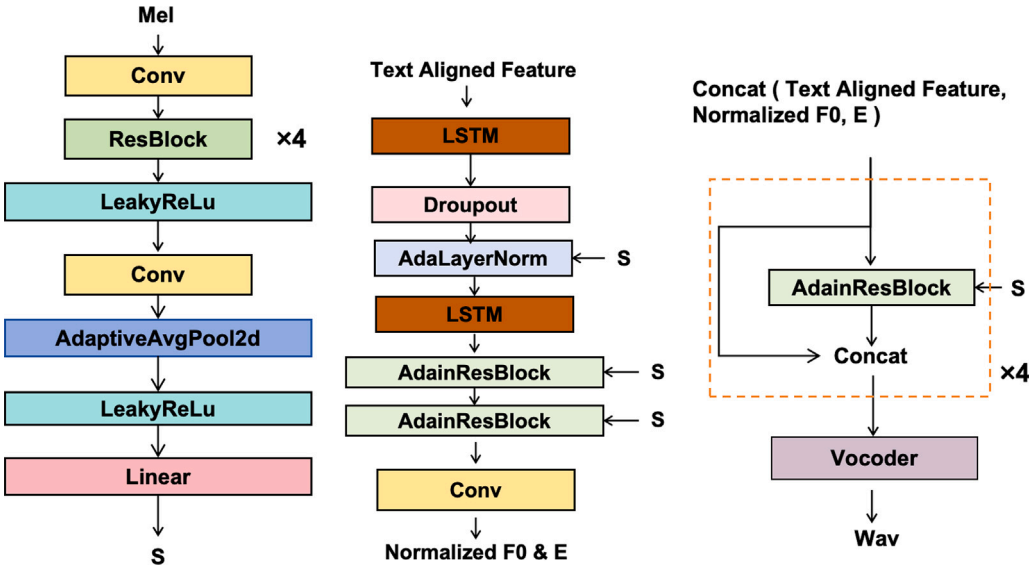
Comprising Speaker Text Encoder and Text Aligner, this module processes phoneme sequences through Conv1D layers, layer normalization, and LSTM to generate phoneme embeddings $\hat{P}$. The Text Aligner we used is AuxiliaryASR.[1] This is a phoneme-level ASR model trained on English speech, which can provide phoneme-level alignment $A^{n\times l}$ for reference speech and phoneme sequences. The model integrates convolutional neural networks with two joint decoders, namely Connectionist Temporal Classification (CTC) and attention-based sequence-to-sequence (S2S) models. This hybrid architecture leverages the training stability of the CTC and the high accuracy of the S2S models, achieving automatic speech recognition at the phoneme level. Temporal-aligned embeddings are computed as:

$$P' = \hat{P} * A \tag{1}$$

3.2.3. Style predictor

The Style Predictor generates normalized F0 $F0_{pred}$, energy E_{pred}, and phoneme durations by fusing text embeddings with prosodic S_p and duration S_d style representations. Normalized F0 preserves rhythmic patterns while reducing speaker-specific characteristics to facilitate style-speaker disentanglement. Energy represents aperiodic components, and phoneme durations determine speech rate - collectively

[1] https://github.com/yl4579/AuxiliaryASR.



(a) Audio Encoder (b) Predictor (c) Decoder

Fig. 3. The internal structures of audio encoder, predictor, and decoder.

defining speech style. The specific structure of this module is shown in the Fig. 3(b).

The duration predictor integrates text embeddings with S_{d} via LSTM with adaptive layer normalization, then generates duration predictions d_pred through a bidirectional LSTM and linear layer. Ground-truth durations d_gt from the Text Aligner provide supervision:

$$L_{dur} = L_1 loss(d_{pred}, d_{gt}) \tag{2}$$

The prosody predictor estimates $F0_{pred}$ and E_{pred} using a shared architecture. Text embeddings and S_{p} are fused through an LSTM with adaptive layer normalization, then aligned with phoneme sequences via matrix multiplication with the alignment matrix A to produce temporal features M_{p}.

Subsequent processing employs two Adaptive Instance Normalization (AdaIN) layers for deep style integration, with Conv1D layers and residual connections enhancing nonlinearity. The final predictions are supervised using z-score normalized ground-truth:

$$L_{normF0} = MSE(F0_{gt}, F0_{pred}) \tag{3}$$

$$L_E = MSE(E_{gt}, E_{pred}) \tag{4}$$

3.2.4. Decoder

The decoder synthesizes the final audio output wav_{syn} from three input components: prosodic features comprising normalized $F0$ and energy content features represented by the aligned phoneme embedding P', and speaker characteristics encoded in the timbre embedding S_t as follows:

$$wav_{syn} = decoder(P', F0_{pred}, N_{pred}, S_t) \tag{5}$$

The decoder's structure is similar to that of HiFiGAN [24]. Unlike the process of recovering waveforms from Mel-spectrogram features, this work focuses on upsampling from normalized $F0$ and other features. To achieve this, we concatenate normalized $F0$, energy E, and phoneme-aligned features along the speech frame duration dimension. After that, we incorporate AdaIN layers in the residual network to fuse speaker-specific features. The fused features are then fed into the vocoder for audio synthesis. We utilize the multi-resolution discriminator (MRD) and the multi-period discriminator (MPD), the same as [25]. The specific structure of this module is shown in the Fig. 3(c).

In our experiments, we found that the generator tends to average high-frequency harmonics, causing a reduction in the speaker's timbre. To address this issue, we implemented multiple sub-discriminators with FFT sizes of 2048 and varying window lengths, enhancing the decoder's sensitivity to high-frequency information. We trained the decoder using a combination of adversarial loss function L_{adv}, feature matching loss L_{fm}, and Mel-spectrogram reconstruction loss L_{mel}, where D represents the two discriminators MPD and MRD. D_i and N_i denote the feature values and the number of features at the i-th layer of the discriminator, respectively. $M_{wav_{syn}}$ represents the computation of the Mel-spectrogram for the synthesized speech.

$$L_{adv}(D,G) = E\left[\left(D\left(wav_{syn}\right)-1\right)^2\right] + E\left[\left(D\left(wav_{syn}\right)\right)^2\right] + E\left[\left(D\left(R\right)\right)^2\right] \tag{6}$$

$$\mathcal{L}_{FM}(G;D) = \mathbb{E}_{(wav_{syn},R)}\left[\sum_{i=1}^{T}\frac{1}{N_i}\left\|D^i(R) - D^i(wav_{syn})\right\|_1\right] \tag{7}$$

$$L_{mel} = L_1 loss\left(M_{spk}, M_{wav_{syn}}\right) \tag{8}$$

3.3. Training process

Our training process is divided into two stages. In the first stage, we only train the speaker encoder, the speaker text encoder in the text adaptor, and the decoder. The training process is shown in Fig. 2, where the modules within the dashed box represent the first phase of training.

The training task at this stage is speech restoration. The speaker encoder extracts the timbre embedding from the speaker reference. The input of the text adaptor is the text corresponding to the speech, and the text is encoded to align with the timbre. In addition, in the first stage, we do not train encoders related to style prediction, but only train encoders to extract timbre. Therefore, in this stage, we extract the normalized $F0$ and Energy E of the speaker reference in advance and directly input the ground truth into the decoder during the synthesis process. Finally, the decoder synthesizes the restored speech, and the synthesized speech is used to calculate the loss with the speaker reference, thus completing

the first stage of training. We use a pre-trained model[2] to extract the normalized $F0$ and Energy E of the speech.

The reason we designed the first stage in this way is that we want to use general speech synthesis tasks to first train an encoder-decoder structure with basic speech synthesis capabilities. This structure will serve as the basis for the second stage of training. By adding encoders that extract duration information and prosody information, respectively, we will eventually obtain a model that can perform controllable speech synthesis tasks. This staged training paradigm enables the gradual increment of model parameters, thereby circumventing redundant computations and enhancing training efficiency. If a single-stage training strategy is employed, the model achieves performance equivalent to that of two-stage training. However, the number of training epochs needed is slightly higher. Specifically, our model undergoes 50 epochs of training in the two-stage paradigm, whereas 60 to 70 epochs are required for the single-stage approach to reach convergence. A similar training process has also been adopted in other related works [21,26].

In the second stage, we jointly train the entire model by adding a prosody encoder, a duration encoder, a style text encoder, and a style predictor, based on the first stage, as shown in Fig. 2. In this stage, normalized $F0$, E, and duration are predicted by the style predictor, and we use the ground truth extracted in the first stage as the supervision signal for these variables. In the training stage, the style reference and speaker reference are the same speech, so we can still use the synthesized speech and speaker reference to calculate the loss as the final loss:

$$L_{final} = L_{dur} + L_{mel} + L_{adv} + L_{FM} + L_{normF0} + L_E \tag{9}$$

4. Experimental setup and evaluation metrics

4.1. Datasets

As shown in Table 2, the datasets we use to train Control-TTS include: LibriTTS [19], ESD [27]. LibriTTS is a multi-speaker English corpus for TTS. We use the two high-quality subsets, train-clean-360 and train-clean-100, as our training set. This part of the data contains 115 h of speech from 1151 speakers. We use dev-clean and test-clean as the validation set and test set, respectively. The ESD (Emotional Speech Database) is a bilingual dataset containing both Chinese and English speech. It includes 350 parallel utterances spoken by 10 native English speakers and 10 native Chinese speakers, covering five emotional categories: neutral, happy, angry, sad, and surprise. The dataset consists of over 29 h of speech recorded in a controlled acoustic environment, designed to support multi-speaker and cross-lingual emotional TTS studies. We use the English part and divide it into the training set, the validation set, and the test set in the ratio of 8:1:1.

During the testing of Control-TTS and other comparative models, we utilized a self-constructed dataset to ensure that the audio within the test dataset had not been encountered by any of the models during their training phases. This part of the data is selected from the VccmDataset proposed by TextrolSpeech [15], with 1000 utterances chosen for WER evaluation and 20 utterances serving as style references and speaker references for subjective evaluation. The selection criteria were clear speech and good audio quality. Additionally, our test cases included some with distinct tones, aiming to assess the models' style cloning capabilities.

4.2. Evaluation metrics

For subjective evaluation, we employed the Naturalness Mean Opinion Score (NMOS) to assess speech quality and naturalness. We employed the Speaker Similarity Mean Opinion Score to assess the timbre similarity with the Speaker Reference and the Style Similarity Mean Opinion Score to evaluate the style similarity with the Style Reference.

[2] https://github.com/yl4579/PitchExtractor

Table 2
Corpus used to train the model.

Corpus	Speech number	Speaker number	Hours
LibriTTS (Train-clean-460)	149,736	1151	115
ESD (English)	17,500	10	15

Both Speaker Similarity and Style Similarity adopt the same human subjective scoring method as NMOS. These three metrics belong to the subjective evaluation category. We selected 15 annotators independent of the model training and development process to conduct blind scoring on speech synthesized by Control-TTS and baseline models. They rated each metric on a scale of 1 to 5, and we calculated the average scores for each model to determine the final scores.

For objective evaluation, we employed the Word Error Rate (WER) to evaluate the clarity of the synthesized speech, reflecting the quality of the synthesized speech. We employed the Whisper [28] speech recognition model to transcribe the test speech, comparing the recognized text with the ground truth text to calculate the WER.

4.3. Model training

We used four NVIDIA A100 GPUs for training and ensured that all training data was resampled to 24 kHz. To maintain data quality and training efficiency, we filtered out speech clips longer than 20 s and shorter than 0.6 s from the dataset. In the first stage, we trained the model for 30 epochs, followed by 20 epochs in the second stage. The training time of each epoch is approximately 5 h. We set the batch size to 32 and used the AdamW optimizer with an initial learning rate of 1e−4. During training, data in each batch were randomly shuffled to ensure that each batch contained different speakers' speech clips.

4.4. Comparison system

Based on the availability of the comparison models, we have selected several TTS systems capable of controllable speech synthesis or accepting two audio inputs as our comparative systems. Among them, PromptStyle [22] and PromptTTS [13] can perform controllable speech synthesis. They use speaker ID to specify the speaker's timbre, and then use text descriptions to describe the voice style. The two together guide the model to synthesize speech. StyleTTS2 [21] supports two speech samples as input, which can be fed into the Prosodic Style Encoder and the Acoustic Style Encoder for encoding, respectively, and then jointly guide speech synthesis. However, StyleTTS2 cannot control the timbre and style separately. The speech received by the two Encoders is the same speech sample, and it can only restore the style of the original voice to a certain extent. Unistyle [17] and ControlSpeech [18] can perform Timbre Clone and Style Control simultaneously, but their styles are controlled by text descriptions.

In terms of model function, both PromptStyle and PromptTTS only support style cloning and cannot clone timbre. Moreover, style control can only be achieved through input text descriptions. StyleTTS2 allows the input of two speech samples, but the original model can only achieve voice cloning when the two speech samples are identical, and it cannot fuse the timbre of one speech sample with the style of another. The styles of Unistyle and ControlSpeech are controlled by text descriptions. In contrast, Control-TTS can simultaneously control both timbre and style using two speech samples and can specify any speech content. This enables our model to mitigate information loss due to text descriptions during the cloning process, clone the speech style better, and enhance control over the speech.

5. Experiment results

5.1. The performance of models across various metrics

In this experiment, we conducted a comprehensive comparison between Control-TTS and various baseline models, evaluating their

Table 3
The performance of the models across various metrics.

Models	NMOS↑	Speaker similarity↑	Style similarity↑	WER↓
PromptStyle	3.58 ± 0.09	–	3.27 ± 0.11	8.1
PromotTTS	3.88 ± 0.15	–	3.29 ± 0.12	4.4
StyleTTS2	4.05 ± 0.08	**3.72 ± 0.15**	3.42 ± 0.09	5.3
UniStyle	3.96 ± 0.12	3.63 ± 0.15	3.43 ± 0.09	5.9
ControlSpeech	4.02 ± 0.11	3.65 ± 0.12	3.40 ± 0.09	4.2
Control-TTS (Ours)	**4.09 ± 0.13**	3.69 ± 0.17	**3.66 ± 0.11**	**3.1**

performance on test audio across several key metrics: Naturalness Mean Opinion Score (NMOS), Speaker Similarity, Style Similarity, and Word Error Rate (WER). To measure speaker similarity, we provided the models with two identical reference speech clips. However, due to the unavailability of training codes for PromptStyle and PromptTTS, we utilized pre-trained versions of these models[3] for testing. These pre-trained versions are fixed-speaker models, and thus, we did not assess their speaker similarity, denoted by "–" in the results table. For the evaluation of style similarity, we employed a consistent speaker reference for voice timbre input while varying the style references. Given that PromptStyle and PromptTTS are designed to accept textual descriptions of speech styles, we substituted audio inputs with corresponding text descriptions. In contrast, StyleTTS2 and Control-TTS were provided with style-embedded audio inputs. This methodological approach ensures a fair and systematic comparison across different models, highlighting the unique capabilities and limitations of each in handling speaker and style variations.

5.1.1. Overall performance

The experimental results are shown in Table 3. The upward arrow indicates that the larger the value is, the better the performance. The downward arrow indicates that the smaller the value is, the better the performance. The ± sign after the score indicates the variance of the model's scores between different test speech samples. The best result for each metric has been highlighted in bold. The experimental results indicate that the Control-TTS model demonstrates outstanding competitiveness across all performance metrics, showing advantages when compared to similar models.

5.1.2. Performance on speech quality and clarity

Specifically, in terms of the MOS for speech naturalness, the Control-TTS model achieved a high score of 4.09, which is significantly better than the 3.58 score of the PromptStyle model and the 3.88 score of the PromptTTS model. This data fully illustrates that, compared to other style-controllable text-to-speech conversion models of the same type, Control-TTS performs more prominently in terms of speech naturalness, with the synthesized speech being closer to natural speech in terms of sound quality, thereby enhancing the auditory experience. Furthermore, the Control-TTS model also surpasses the traditional speech synthesis model StyleTTS2 with a score of 4.05 in terms of naturalness, further confirming the significant progress made by the Control-TTS model in sound quality.

In the assessment of speech quality, in addition to subjective evaluations, objective measurement indicators also play a crucial role. Among these, the WER is an important quantitative metric used to measure the accuracy of speech recognition. The experimental data show that the Control-TTS model significantly outperforms other comparative models with a WER of 3.1, reflecting its exceptional performance in speech clarity. This indicates that the Control-TTS model synthesizes audio with higher clarity, exhibiting fewer misreadings and pronunciation ambiguities compared to other models. This advantage not only highlights its superior performance in speech synthesis tasks but also provides robust support for its practical application in multimodal human-computer interaction.

5.1.3. Performance on style similarity and speaker similarity

In terms of style similarity, Control-TTS outperformed other models. Compared with models that rely on text descriptions as style guidance, Control-TTS adopts a more direct approach, using style speech as input prompts to more accurately extract the style features of the reference speech. This approach mitigates the errors and information loss that may occur in the text-to-speech modality conversion process.

When users listen to the stylized speech synthesized by Control-TTS, they can more clearly perceive the model's cloning of subtle changes in the reference speech, such as pauses, haste, and intonation. In contrast, the PromptStyle and PromoteTTS models based on text descriptions can only achieve a rough simulation of the speech style and cannot align with the original speech at a fine-grained level. Control-TTS directly extracts information from the style reference speech, preventing the loss of stylistic details. In contrast, PromptStyle and PromptTTS rely on textual prompts to describe audio styles, which often fail to comprehensively capture the nuances present in the reference speech.

In addition, Control-TTS's performance in the style cloning task also surpassed the StyleTTS2 model that uses voice cloning technology. This result further confirms the applicability and accuracy of using dual-style encoders to model style in style transfer. In the subsequent experimental section, we will conduct a more in-depth analysis and verification of the style modeling method, aiming to provide sufficient evidence to support the effectiveness of our method.

In terms of speaker similarity evaluation, Control-TTS shows comparable performance to the specialized voice cloning model StyleTTS2. This result is crucial because it shows that Control-TTS does not negatively affect the speaker timbre cloning effect during style modeling. In other words, our model not only maintains speaker timbre replication capabilities comparable to those of models such as StyleTTS2 but is also able to inject richer and more diverse style features while retaining the original speaker timbre. This finding reveals the flexibility of the Control-TTS model in multitasking, that is, it can achieve fine-tuning and control of style while performing voice cloning tasks. This capability not only expands the application scope of the model, making it suitable for scenarios that require accurate speaker imitation, but also enhances the model's expressiveness in speech synthesis, allowing users to customize the style and emotional expression of the voice as required.

5.2. The comparison of performance using different numbers of style encoders

In Control-TTS, we use two encoders to model speech style collaboratively. The prosody encoder and prosody predictor are responsible for extracting and predicting the normalized F0 and energy N of the style reference speech. Meanwhile, the duration encoder and duration predictor focus on extracting and predicting the duration of each phoneme in the style reference speech. This dual-encoder approach reduces the encoding load on any single encoder, thereby avoiding inaccuracies in encoding and prediction that might arise from handling multiple training objectives simultaneously. This approach enhances the accuracy of variable predictions and improves the effectiveness of style cloning.

In this experiment, we compare Control-TTS with two style encoders, as used in practical applications, against a version with only one style encoder. In the version with only one style encoder, the only style encoder is responsible for encoding normalized $F0$, E, and the duration of each phoneme at the same time, instead of utilizing two encoders as in the full Control-TTS. Through this experiment, we aim to demonstrate the effectiveness of using dual encoders for speech style modeling, thereby further validating our approach. The experimental results are shown in Table 4.

In our evaluation across three key performance metrics, the Control-TTS model with dual style encoders demonstrated notable

[3] https://github.com/yl4579/PitchExtractor.

Table 4
The comparison of performance using different numbers of style encoders.

Model	NMOS↑	Speaker similarity↑	Style similarity↑
Control-TTS (Only one style encoder)	3.85 ± 0.13	3.65 ± 0.17	2.89 ± 0.08
Control-TTS (Prosody encoder + Duration encoder)	**4.09 ± 0.13**	**3.69 ± 0.17**	**3.66 ± 0.11**

superiority, achieving higher naturalness (NMOS) scores and greater speaker similarity. Particularly noteworthy is that the dual-encoder version of Control-TTS showed a significant advantage in style similarity, substantially outperforming the single-encoder version. The dual-encoder model was able to more precisely reproduce intonation variations and speech rate, resulting in generated speech that closely reflected the original style in the style transfer process.

In practical tests, we observed that the single-encoder model could only partially replicate the target speech rate during style transfer, while its tone presentation tended to be relatively flat, almost devoid of expressive intonation. The dual-encoder model, on the other hand, achieved a marked improvement in tone fidelity, closely mirroring the emotional expression and intonational variation of the source style. Our analysis suggests that this performance difference arises from the single encoder's difficulty in simultaneously encoding prosodic features and the duration information of each phoneme. This limitation leads the single-encoder model to predict speech duration (Duration) with a degree of accuracy but struggle to accurately predict Normalized fundamental frequency (Normalized F0), resulting in inadequate tone reproduction.

In contrast, the dual-encoder model exhibits a more effective division of tasks, with each encoder specializing in the prediction of either Normalized F0 or duration. This focused approach allows the model to maximize its strengths, resulting in an optimal balance between tone, prosody, and speech rate reproduction. The dual-encoder architecture, through task-specific specialization, significantly enhances model performance in style transfer, generating speech that is both more natural and lifelike, successfully capturing the emotional and stylistic elements of the source speech.

5.3. T-SNE clustering experiment

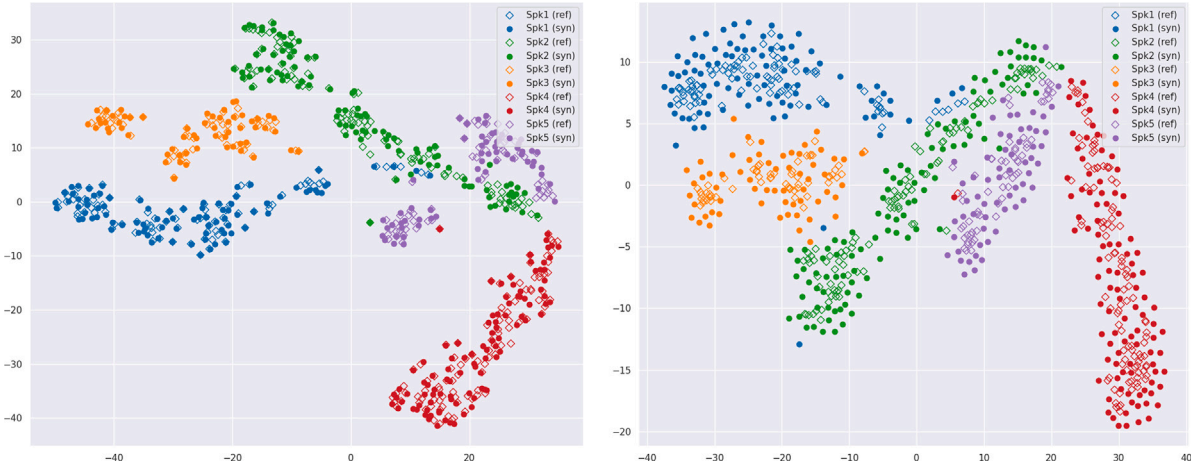
In this experiment, we employ the t-SNE clustering method to evaluate the capability of Control-TTS in extracting timbre and style, aiming to ascertain whether Control-TTS can effectively differentiate between various speakers and speaking styles, which is subsequently reflected in the synthesized audio. We selected a test dataset from the portion of the ESD dataset not utilized for training, which comprises utterances from five distinct speakers, each expressing different emotions, including sadness, neutrality, and anger. Each audio sample from the test dataset was individually input into the model as a reference speech, prompting the model to clone the timbre and style of the reference speech, culminating in the output of synthesized audio. Our experiment was bifurcated into two segments: the first scrutinized the model's proficiency in distinguishing between the timbres of different speakers, while the second assessed its ability to differentiate between various speaking styles.

5.3.1. Timbre extraction experiment

In the timbre extraction experiment, we first input the test audio into the speaker encoder of Control-TTS to extract the speaker embedding from the reference speech. Subsequently, we use the audio from the test set as a reference to generate corresponding synthetic audio via Control-TTS. The synthesized audio is then fed back into the speaker encoder of Control-TTS to extract its speaker embedding. To analyze these embeddings, we employ the t-SNE method to reduce the dimensionality of the speaker embeddings and visualize the results in a two-dimensional coordinate system. By observing the distribution of data points in the two-dimensional space, we can evaluate the effectiveness of Control-TTS in distinguishing different speakers and assess whether the timbre characteristics of the reference speech are preserved during the synthesis process. We test two versions of Control-TTS separately: one with both a Prosody encoder and a Duration encoder, and the other with only one style encoder. The test results of both models are plotted in the Fig. 4.

The experimental results are illustrated in Fig. 4, where distinct colors are utilized to differentiate between speakers. For each speaker, the reference speech is denoted by hollow diamonds, whereas the synthesized audio is represented by solid circles. Among the five speakers, speakers 1, 2, and 3 are male, while speakers 4 and 5 are female. In the t-SNE visualization, the x-axis and y-axis do not carry any specific meaning but serve as coordinate representations of the vectors. The primary objective is to analyze the relative spatial relationships between the vectors.

As observed from Fig. 4, for Control-TTS with two style encoders, data points of different colors exhibit a pronounced clustering trend in the space, with relatively clear boundaries between clusters and



(a) Prosody encoder + Duration encoder (b) Single style encoder

Fig. 4. The speaker encoder of Control-TTS generates embeddings for different speakers, where each point represents a segment of audio, and points of different colors denote distinct speakers. On the left is the clustering result of Control-TTS with both a Prosody encoder and a Duration encoder, and on the right is the clustering result of Control-TTS with only one style encoder.

(a) Control-TTS (b) StyleTTS2

Fig. 5. T-SNE experiment on style extraction. The emotion embeddings extracted from the audio synthesized by Control-TTS and StyleTTS2 are represented as points, where each point corresponds to a segment of synthesized audio, and points of different colors signify distinct styles. Control-TTS exhibits distinct emotional expression across various style reference speeches, whereas StyleTTS2 fails to differentiate the three emotions effectively, resulting in a blend of emotional traits.

no intersections among data points of different colors. This indicates that Control-TTS is capable of effectively distinguishing between the timbres of different speakers, encoding their audio into independent, non-overlapping information, thereby avoiding confusion among the timbres of different speakers. Additionally, we observe that the three clusters representing male voices are predominantly located on the left side of the figure, while the two clusters representing female voices are mainly distributed on the right. This phenomenon suggests that Control-TTS not only differentiates between the timbral characteristics of male and female voices, but also that this differentiation is based on specific timbral patterns rather than random separation. Consequently, through this experiment, we have visually validated the efficacy of Control-TTS in timbre extraction and differentiation, further supporting its potential in multi-speaker voice processing applications. In the case of having only one style encoder, the clustering boundaries between different speakers are not as clear, and the speaker vectors exhibit a certain degree of deviation.

Furthermore, it is observed that the speaker embeddings of different speakers show negligible variations before and after synthesis, maintaining their inherent clustering properties. As illustrated in Fig. 4, the distributions of circles (synthesized audio) and diamonds (reference speech) of the same color remain closely aligned. This demonstrates that Control-TTS successfully preserves the timbre of the reference speech during the synthesis process, ensuring that the speaker's voice characteristics remain consistent. In the case of having only one style encoder, the degree of deviation of the synthesized speech compared to the reference speech is slightly larger.

5.3.2. Style extraction experiment

In the style extraction experiment, we utilized an additional tool[4] for extracting emotion embeddings to derive emotion embeddings from the synthesized audio. Subsequently, we applied the t-SNE method to reduce the dimensionality of these emotion embeddings and performed clustering, plotting the results on a two-dimensional coordinate system. By examining the distribution of data points within this two-dimensional space, we aimed to evaluate the ability of Control-TTS to effectively distinguish between different speaking styles. In a similar vein, we also assessed the capability of the baseline model, StyleTTS2, to differentiate among various speaking styles, and we plotted these results alongside those of Control-TTS in the same two-dimensional coordinate system for comparative analysis.

[4] https://github.com/ddlBoJack/emotion2vec.

Fig. 5 illustrates the synthesized results of Control-TTS for different emotional reference speeches, where the emotions of sad, neutral, and angry are distinctly clustered in three separate directions, demonstrating a clear degree of separation. Although a degree of overlap is observed at the boundaries between these emotional categories, Control-TTS significantly outperforms the baseline model, StyleTTS2, in terms of emotional distinction. As shown in the right panel of the figure, StyleTTS2 fails to differentiate between the three emotions, resulting in a complete blending of emotional features during synthesis. In contrast, our proposed Control-TTS effectively preserves the distinctions among these emotions, highlighting its superior capability in emotion-aware speech synthesis.

5.4. Experiment on the integration of timbre and style

In the scenario where the model is provided with two distinct audio inputs for the integration of timbre and style, a critical requirement is the model's ability to independently extract the timbre from the speaker reference speech and the style from the style reference speech, without conflating the timbres or styles of the two reference speeches. Consequently, in this experiment, each test case we selected comprises two different audio samples. For the speaker similarity metric, we concurrently examine the synthesized audio's similarity to both the speaker reference and the style reference in terms of speaker characteristics. Similarly, for the style similarity metric, we assess the synthesized audio's similarity to both references in terms of stylistic elements. It is only when a discernible gap in similarity to the two references is observed for the same metric that we can affirm the model's correct extraction of the timbre from the speaker reference speech and the style from the style reference speech, rather than confusion between the two. In this experiment, we have chosen StyleTTS2 as the baseline model for comparison. We input two distinct reference speeches into StyleTTS2's Prosodic Style Encoder and Acoustic Style Encoder, respectively.

The experimental results are presented in Table 5, where each metric is accompanied by an arrow indicating the desired direction of the value—upward arrows denote that higher values are preferable, while

Table 5
The performance of the models on the integration of timbre and style.

Model	Speaker Similarity with		Style Similarity with	
	Speaker Reference↑	Style Reference↓	Speaker Reference↓	Style Reference↑
StyleTTS2	3.25	0.95	2.64	2.93
Control-TTS	**4.25**	**0.58**	**2.05**	**3.53**

downward arrows signify that lower values are better. As evidenced by the data in the table, Control-TTS exhibits a higher degree of speaker similarity to the Speaker Reference and a lower degree to the Style Reference, outperforming StyleTTS2 in both speaker similarity metrics. Compared to StyleTTS2, Control-TTS more distinctly differentiates the timbres of the two references, with the synthesized audio's timbre being closer to that of the Speaker Reference. Similarly, Control-TTS achieves a higher style similarity to the Style Reference and a lower similarity to the Speaker Reference, excelling over StyleTTS2 in the independent extraction of the Style Reference's style. These experimental outcomes demonstrate that the Control-TTS model accurately extracts the timbre from the speaker reference speech and the style from the style reference speech, without conflating the timbres or styles of the two reference speeches.

5.5. The impact of reference speech length on synthesis performance

To investigate the impact of reference speech length on the performance of synthesized speech, we collected four categories of reference speech with different durations: less than 2s, 2s-5s, 5s-10s, and 10s-20s. We then evaluated the Mean Opinion Score (MOS) for speech quality, speaker similarity, and style similarity under these four conditions, with the experimental results illustrated in the line chart. In the design of the proposed model, zero-padding is applied to all speech segments shorter than 0.6s. The experimental results show that when the duration of the reference audio is less than 2 s, the quality of the synthesized audio decreases slightly, and when the duration reaches more than 2 s, the quality of the synthesized audio tends to stabilize. Notably, longer reference speech durations can slightly improve speaker similarity, indicating that extended speech segments can provide more accurate speaker representations (Fig. 6).

5.6. The impact of noisy reference audio on the synthesis performance

To verify the model's robustness to noisy speech inputs, we designed three types of noisy speech input experiments: specifically, experiments where the speaker reference speech contains noise, where the style reference speech contains noise, and where both reference speeches contain noise. The experimental results are presented in Table 6. When the speaker reference speech is noisy, the quality of the synthesized speech decreases slightly. This is because the speaker encoder interprets background noise as one of the speaker's style characteristics, causing the synthesized speech to replicate the noise and thus lower the Mean Opinion Score (MOS) for quality, though it barely affects style similarity. In contrast, the style encoder shows better robustness to noisy speech; this is hypothesized to stem from the theoretical constraints of normalized fundamental frequency and energy, which guide the style encoder to ignore part of the noise, resulting in only a slight decrease in both the quality and similarity of the synthesized speech.

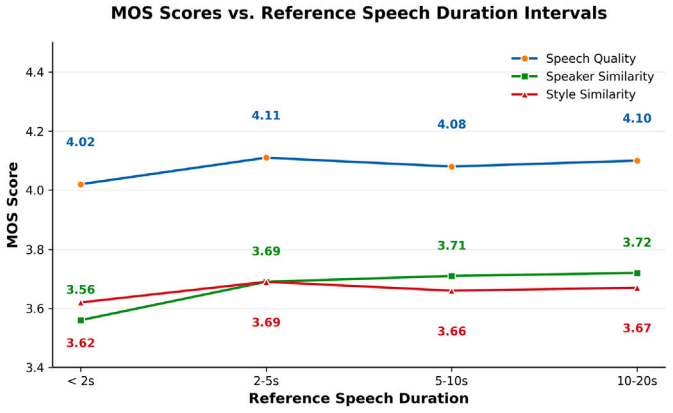


Fig. 6. The impact of reference speech length on synthesis performance.

Table 6
The impact of Chinese reference audio on synthesis performance.

Noise addition method	NMOS↑	Speaker Similarity↑	Style Similarity↑	WER↓
Speaker Refernce	3.74 ± 0.04	3.45 ± 0.02	3.53 ± 0.04	7.8
Style Reference	3.90 ± 0.04	3.56 ± 0.03	3.48 ± 0.06	7.8
Both	3.71 ± 0.04	3.43 ± 0.05	3.42 ± 0.09	8.9
None	**4.09 ± 0.13**	**3.69 ± 0.17**	**3.66 ± 0.11**	**3.1**

Table 7
The impact of the amount of training data on synthesis performance.

Model	NMOS↑	Speaker Similarity↑	Style Similarity↑	WER↓
Half training data	3.73 ± 0.11	3.52 ± 0.09	3.35 ± 0.12	5.2
Original model	**4.09 ± 0.13**	**3.69 ± 0.17**	**3.66 ± 0.11**	**3.1**

Table 8
The impact of Chinese reference audio on synthesis performance.

Speaker + Style	NMOS↑	Speaker Similarity↑	Style Similarity↑	WER↓
ch + ch	3.94 ± 0.07	3.47 ± 0.05	3.59 ± 0.13	4.6
ch + en	4.02 ± 0.13	3.56 ± 0.02	3.61 ± 0.06	4.4
en + ch	4.03 ± 0.05	3.63 ± 0.10	3.64 ± 0.06	3.6
en + en	**4.09 ± 0.13**	**3.69 ± 0.17**	**3.66 ± 0.11**	**3.1**

5.7. The impact of the amount of training data on synthesis performance

In this experiment, we randomly sampled half of the original data to train a new model. By comparing the differences in synthesis performance between the new model and the original model, we explored the impact of the amount of training data on the model's synthesis performance. Our test results are shown in the Table 7. After reducing the amount of training data, the model's performance on various metrics decreased slightly compared to the original model. This indicates that a larger amount of training data helps to improve the model's synthesis performance.

5.8. The impact of Chinese reference audio on synthesis performance

To verify the model's generalization for cross-lingual reference speech input tasks, experiments were conducted using Chinese data from the ESD dataset, with results summarized in Table 8. Four comparative experimental groups were designed: "ch + ch" denotes both speaker reference and style reference speech are in Chinese; "ch + en" indicates a Chinese speaker reference paired with an English style reference; "en + ch" represents an English speaker reference combined with a Chinese style reference.

It was observed that using Chinese speech as the speaker reference leads to a slight decrease in speaker similarity, because the pronunciation rules of different languages cause varying degrees of changes in timbre. Meanwhile, whether Chinese or English is used as the emotional reference, the style similarity and NMOS score of the synthesized speech only show slight and comparable decreases. This indicates that the decoder and style speech encoder in the proposed system have strong cross-lingual generalization capabilities.

6. Conclusion

In the field of multimodal human-computer interaction, the demand for customizable speech among users has become increasingly prominent, especially in terms of precise control over timbre and style. To address this core requirement, this paper proposes a novel

task: Controllable Timbre Cloning and Style Replication with Reference Speech Examples, and designs the Control-TTS model as a solution. The model directly controls the speaker's timbre and speaking style of the synthesized speech through reference speech, extracting the timbre from the speaker's reference speech and the prosody from the speaking style reference speech, respectively, thereby achieving the free combination of the two. This effectively enhances the diversity of speech generation, providing a more personalized speech output solution for multimodal human-computer interaction scenarios.

To improve the accuracy of style replication, this paper innovatively adopts a multi-encoder architecture to model speech style from multiple dimensions, such as prosody and speaking rate. This multi-perspective style feature capture mechanism can more comprehensively capture style details in the reference speech, thus enabling faithful restoration of the original style during generation.

Experimental results on the VccmDataset fully verify the effectiveness of Control-TTS, as it achieves comparable or state-of-the-art performance in key metrics, including naturalness mean opinion score (NMOS), word error rate (WER), speaker similarity, and style similarity. Meanwhile, the experiments further confirm that the strategy of modeling speech style from different perspectives through multiple encoders significantly improves the cloning accuracy of style information, providing strong support for achieving higher-quality style replication. In summary, through the innovative task definition and model design, Control-TTS breaks through the limitations of existing technologies in the joint control of timbre and style, provides a new technical path for personalized speech in multimodal human-computer interaction, and is expected to promote the development of more natural and flexible multimodal human-computer interaction systems.

CRediT authorship contribution statement

Tianwei Lan: Writing – original draft, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Yuhang Guo:** Methodology, Investigation, Formal analysis, Conceptualization. **Mengyuan Deng:** Methodology, Formal analysis, Data curation, Conceptualization. **Jing Wang:** Investigation, Funding acquisition, Conceptualization. **Wenwu Wang:** Writing – original draft, Investigation. **Chong Feng:** Investigation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Jing Wang reports that financial support was provided by the Beijing Natural Science Foundation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by National Natural Science Foundation of China (Nos. 62571037, U25B2075), Beijing Natural Science Foundation (Nos. L242089, L257001).

Data availability

Data will be made available on request.

References

[1] B. Yu, S. Zhang, L. Zhou, J. Wei, L. Sun, L. Bu, Brain-inspired computing based on deep learning for human-computer interaction: a review, Neurocomputing 650 (2025) 130928, https://doi.org/10.1016/j.neucom.2025.130928, https://www.sciencedirect.com/science/article/pii/S0925231225016005.

[2] Y. Wu, X. Wang, D. Li, R. Hu, Multimodal and multichannel speech separation using location-guided speech feature mapping network, Neurocomputing 652 (2025) 131051, https://doi.org/10.1016/j.neucom.2025.131051, https://www.sciencedirect.com/science/article/pii/S0925231225017230.

[3] X. Qi, Y. Wen, P. Zhang, H. Huang, MFGCN: multimodal fusion graph convolutional network for speech emotion recognition, Neurocomputing 611 (2025) 128646, https://doi.org/10.1016/j.neucom.2024.128646, https://www.sciencedirect.com/science/article/pii/S0925231224014176.

[4] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, F. Wei, Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers, arXiv e-prints arXiv:2301.02111, 2023.

[5] K. Shen, Z. Ju, X. Tan, Y. Liu, Y. Leng, L. He, T. Qin, S. Zhao, J. Bian, NaturalSpeech 2: Latent Diffusion Models are Natural and Zero-Shot Speech and Singing Synthesizers, arXiv e-prints arXiv:2304.09116, 2023.

[6] Z. Jiang, Y. Ren, Z. Ye, J. Liu, C. Zhang, Q. Yang, S. Ji, R. Huang, C. Wang, X. Yin, Z. Ma, Z. Zhao, Mega-TTS: Zero-Shot Text-to-Speech at Scale with Intrinsic Inductive Bias, arXiv e-prints arXiv:2306.03509, 2023.

[7] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, W.-N. Hsu, Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale, arXiv e-prints arXiv:2306.15687, 2023.

[8] J. Li, L. Zhang, Zse-Vits: a zero-shot expressive voice cloning method based on Vits, Electronics 12 (2023) 820.

[9] Y. Jiang, T. Li, F. Yang, L. Xie, M. Meng, Y. Wang, Towards Expressive Zero-Shot Speech Synthesis with Hierarchical Prosody Modeling, arXiv e-prints arXiv:2406.05681, 2024.

[10] H. Tang, X. Zhang, J. Wang, N. Cheng, J. Xiao, EmoMix: Emotion Mixing via Diffusion Models for Emotional Speech Synthesis, arXiv e-prints arXiv:2306.00648, 2023.

[11] Z. Li, X. Xing, J. Wang, S. Chen, G. Yu, G. Wan, X. Xu, AS-Speech: Adaptive Style For Speech Synthesis, arXiv e-prints arXiv:2409.05730, 2024.

[12] X. Zhu, Y. Lei, T. Li, Y. Zhang, H. Zhou, H. Lu, L. Xie, METTS: Multilingual Emotional Text-to-Speech by Cross-speaker and Cross-lingual Emotion Transfer, arXiv e-prints arXiv:2307.15951, 2023.

[13] Z. Guo, Y. Leng, Y. Wu, S. Zhao, X. Tan, PromptTTS: Controllable Text-to-Speech with Text Descriptions, arXiv e-prints arXiv:2211.12171, 2022.

[14] D. Yang, S. Liu, R. Huang, C. Weng, H. Meng, InstructTTS: Modelling Expressive TTS in Discrete Latent Space with Natural Language Style Prompt, arXiv e-prints arXiv:2301.13662, 2023.

[15] S. Ji, J. Zuo, M. Fang, Z. Jiang, F. Chen, X. Duan, B. Huai, Z. Zhao, TextrolSpeech: A Text Style Control Speech Corpus With Codec Language Text-to-Speech Models, arXiv e-prints arXiv:2308.14430, 2023.

[16] T. Li, C. Hu, J. Cong, X. Zhu, J. Li, Q. Tian, Y. Wang, L. Xie, DiCLET-TTS: Diffusion Model based Cross-lingual Emotion Transfer for Text-to-Speech – A Study between English and Mandarin, arXiv e-prints arXiv:2309.00883, 2023.

[17] X. Zhu, W. Tian, X. Wang, L. He, Y. Xiao, X. Wang, X. Tan, S. Zhao, L. Xie, Unistyle: unified style modeling for speaking style captioning and stylistic speech synthesis, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 7513–7522.

[18] S. Ji, J. Zuo, W. Wang, M. Fang, S. Zheng, Q. Chen, Z. Jiang, H. Huang, Z. Wang, X. Cheng, et al., Controlspeech: Towards simultaneous zero-shot speaker cloning and zero-shot language style control with decoupled codec, arXiv preprint arXiv:2406.01205, 2024.

[19] H. Zen, V. Dang, R. Clark, Y. Zhang, R.J. Weiss, Y. Jia, Z. Chen, Y. Wu, LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech, arXiv e-prints arXiv:1904.02882, 2019.

[20] Y. Leng, Z. Guo, K. Shen, X. Tan, Z. Ju, Y. Liu, Y. Liu, D. Yang, L. Zhang, K. Song, L. He, X.-Y. Li, S. Zhao, T. Qin, J. Bian, PromptTTS 2: Describing and Generating Voices with Text Prompt, arXiv e-prints arXiv:2309.02285, 2023.

[21] Y.A. Li, C. Han, V.S. Raghavan, G. Mischler, N. Mesgarani, StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models, arXiv e-prints arXiv:2306.07691, 2023.

[22] G. Liu, Y. Zhang, Y. Lei, Y. Chen, R. Wang, Z. Li, L. Xie, PromptStyle: Controllable Style Transfer for Text-to-Speech with Natural Language Descriptions, arXiv e-prints arXiv:2305.19522, 2023.

[23] Y.A. Li, C. Han, X. Jiang, N. Mesgarani, Phoneme-Level BERT for Enhanced Prosody of Text-to-Speech with Grapheme Predictions, arXiv e-prints arXiv:2301.08810, 2023.

[24] J. Kong, J. Kim, J. Bae, HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis, arXiv e-prints arXiv:2010.05646, 2020.

[25] S.-G. Lee, W. Ping, B. Ginsburg, B. Catanzaro, S. Yoon, Bigvgan: A universal neural vocoder with large-scale training, arXiv preprint arXiv:2206.04658, 2022.

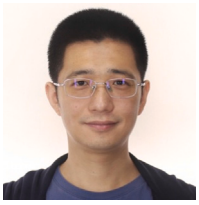
[26] Y.A. Li, C. Han, N. Mesgarani, Styletts: a style-based generative model for natural and diverse text-to-speech synthesis, IEEE J. Sel. Top. Signal Process. 19 (2025) 283–296, https://doi.org/10.1109/JSTSP.2025.3530171

[27] K. Zhou, B. Sisman, R. Liu, H. Li, Emotional Voice Conversion: Theory, Databases and ESD, arXiv e-prints arXiv:2105.14762, 2021.

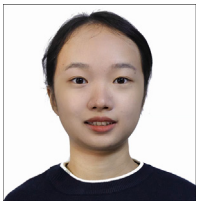
[28] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust Speech Recognition via Large-Scale Weak Supervision, arXiv e-prints arXiv:2212.04356, 2022.

Author biography

Tianwei Lan is a master's student specializing in natural language processing. His research interests include speech synthesis, information extraction, and language model security.



Yuhang Guo is a lecturer with research interests in natural language processing, information extraction, machine translation, machine learning, and artificial intelligence.



Mengyuan Deng is a master's student. Her research interests include speech deepfake detection and speech synthesis.

Jing Wang received the Ph.D. degree in communication and information system from Beijing Institute of Technology (BIT) in China in 2007. She works as an Associate Professor at the Research Institute of Communication Technology (RICT), School of Information and Electronics, Beijing Institute of Technology. Her current research is speech and audio signal processing, multimedia communication and virtual reality. She is IEEE Senior Member, AES Member, Senior Member of CIC (China Institute of Communications), Senior Member of CIE (Chinese Institute of Electronics), Member of CCF (China Computer Federation), and an expert member of Digital Audio and Video Coding Standard Workgroup (AVS) in China.

Wenwu Wang (Senior Member, IEEE) was born in Anhui, China. He received the B.Sc., M.E., and Ph.D. degrees from Harbin Engineering University, Harbin, China, in 1997, 2000, and 2002, respectively. He then worked with King's College London, London, U.K., Cardiff University, Cardiff, U.K., Tao Group Ltd.(now Antix Labs Ltd.), Reading, U.K., and Creative Labs, before joining the University of Surrey, Guildford, U.K., in 2007, where he is currently a Professor of signal processing and machine learning. He is also an AI Fellow with the Surrey Institute for People Centred Artificial Intelligence. His research interests include signal processing, machine learning and perception, artificial intelligence, machine audition (listening), and statistical anomaly detection. He has coauthored more than 300 papers in these areas. He has been recognized as a coauthor or corecipient of more than 15 awards, including the 2022 IEEE Signal Processing Society Young Author Best Paper Award, ICAUS 2021 Best Paper Award, DCASE 2020 and 2023 Judge's Award, DCASE 2019 and 2020 Reproducible System Award, and LVA/ICA 2018 Best Student Paper Award. He is an Associate Editor (2020–2025) for IEEE/ACM TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING and Associate Editor (2024–2026) for IEEE TRANSACTIONS ON MULTIMEDIA. He was a Senior Area Editor (2019–2023) and Associate Editor (2014–2018) for IEEE TRANSACTIONS ON SIGNAL PROCESSING. He is the elected Chair (2023–2024) of IEEE Signal Processing Society (SPS) Machine Learning for Signal Processing Technical Committee, Board Member (2023–2024) of IEEE SPS Technical Directions Board, elected Chair (2025–2027) and Vice Chair (2022–2024) of the EURASIP Technical Area Committee on Acoustic Speech and Music Signal Processing, elected Member (2021–2026) of the IEEE SPS Signal Processing Theory and Methods Technical Committee. He was a Satel-lite Workshop Co-Chair for IEEE ICASSP 2024 and INTERSPEECH 2022, Publication Co-Chair for IEEE ICASSP 2019, Local Arrangement Co-Chair of IEEE MLSP 2013, and Publicity Co-Chair of IEEE SSP 2009. He is a Special Session Co-Chair of IEEE MLSP 2024, and Technical Program Co-Chair of IEEE MLSP 2025.

Chong Feng received his PhD degree from University of Science and Technology of China in 2005. He is currently a professor at School of Computer Science & Technology, Beijing Institute of Technology, China, where he also serves as the vice director of Language Intelligence and Social Computing Research Center. Prof. Feng has published numerous papers in international journals and conferences, such as TKDE, KBS, SIGIR, EMNLP, AAAI. He has served as a reviewer for international SCI journals such as KBS and TKDE, and as an Area Chair and Program Committee Member for various domestic and international conferences, including COLING, CCL, and CCMT. His research interests include social media, knowledge graph, machine translation, natural language processing and large languuge model.