Controllable Text-to-Speech Synthesis with Masked-Autoencoded Style-Rich Representation

Anonymous ACL submission

Abstract

Controllable text-to-speech (TTS) systems aim to manipulate various stylistic attributes of generated speech. Existing models that use natural language prompts as an interface often lack the ability for fine-grained control and face a scarcity of high-quality data. To address these challenges, we propose a two-stage style-controllable TTS system with language models, utilizing a masked-autoencoded representation as an intermediary. We employ a masked autoencoder to learn a speech feature rich in stylistic information, which is then discretized using a residual vector quantizer. In the first stage, an autoregressive transformer is used for the conditional generation of these style-rich tokens from text and control signals. In the second stage, we generate codec tokens from both text and sampled style-rich tokens. Experiments demonstrate that training the first-stage model on extensive datasets enhances the robustness of the two-stage model in terms of quality and content accuracy. Additionally, our model achieves superior control over attributes such as pitch and emotion. By selectively combining discrete labels and speaker embeddings, we can fully control the speaker's timbre and other stylistic information, or adjust attributes like emotion for a specified speaker. Audio samples are available at https://style-ar-tts.github.io.

1 Introduction

Controllable text-to-speech (TTS) systems aim to generate high-fidelity speech while allowing control over various style attributes of the synthesized speech, such as speaker timbre, pitch level and variation, emotion, acoustic environment, etc. Due to its promising applications in digital media production and human-computer interaction, controllable TTS has been attracting growing interest in the machine learning community with a substantial amount of research working on it (Guo et al., 2023; Leng et al., 2023; Ji et al., 2024; Yang et al., 2024; Zhou et al., 2024).

Despite the extensive research on this topic, controllable TTS still faces some unsolved challenges: **(1) Control Interface Issue**. Most existing works use natural language prompts as a medium of style control, which is friendly for non-professional users. However, style descriptions with natural language tend to be broad and coarse-grained, making it difficult to precisely control specific attributes. Moreover, the rich diversity of natural language brings more challenges to modeling the relationship between style attributes and prompts. It is also difficult to fully encompass the user instructions in real-world scenarios, restricting the application of these methods. **(2) Data Issue**. The training of well-performed TTS systems relies on high-quality speech corpora, which are often limited in both data volume and stylistic diversity. When using natural language as the control interface, the additional cost of generating prompt sentences further restricts the data size. Present controllable TTS datasets (Guo et al., 2023; Ji et al., 2024) are often limited to hundreds of hours. This constraint puts challenges on learning precise control abilities and improving generation diversity.

In this paper, we propose a fine-grained controllable TTS system. In contrast to natural language prompts, We divide the value ranges of various stylistic attributes of speech into multiple intervals, each represented by a label, and use these labels as conditional inputs to achieve fine-grained control. By selectively combining these labels with speaker embeddings, we can generate new speaker timbre while controlling other attributes, or adjust certain attributes such as emotion for a given speaker.

Our controllable TTS system adopts a two-stage generation paradigm using two language models (LMs), with a style-rich representation as an intermediate output. We adopt a masked autoencoder (MAE) which learns to capture diverse style information by reconstructing mel filterbank from the encoded content input and masked fbank. The features extracted by the style encoder of the trained MAE are then discretized and used as an intermediary of the TTS pipeline. Each of the two stages relies on a decoder-only transformer. The first stage generates style-rich tokens conditioned on content and control signals, while the second stage generates codec tokens from the content input and the predicted style-rich tokens. Due to low dependence on high-quality corpora, the style-rich token generation phase can scale up to a large amount of data, boosting control capability and generation diversity; while in the codec token generation stage, a relatively small amount of data is sufficient to learn how to reconstruct codec tokens from content and style units, addressing the issue of high-quality data scarcity. To enhance the control accuracy of fine-grained attributes, we investigate classifier-free guidance in the style-rich token generation stage. Experiments indicate that our model achieves good style control ability while keeping decent audio quality and content accuracy.

2 Related works

2.1 Controllable text-to-speech

Controllable TTS aims to enable control over stylistic attributes of the speech during synthesis. The earliest exploration, PromptTTS (Guo et al., 2023), extracts textual features from prompts with a fine-tuned BERT and incorporates them in a TTS backbone with attention. InstructTTS (Yang et al., 2024) achieves a text-controlled expressive TTS system with cross-modal representation learning. PromptTTS 2 (Leng et al., 2023) employs a variational network to generate reference acoustic features conditioned on text features. Audiobox (Vyas et al., 2023) builds a unified natural-language-instructed flow-matching model integrating speech, music, and audio generation. TextrolSpeech (Ji et al., 2024) integrates natural language style prompt into the condition of VALL-E (Wang et al., 2023a) for controllable TTS. VoxInstruct (Zhou et al., 2024) merges the content input and style prompt of TTS into a single composite textual instruction and utilizes a multimodal codec language model as the backbone for TTS. Unlike these methods using natural language as the control interface, we adopt a two-stage controllable TTS system with attribute labels for fine-grained control.

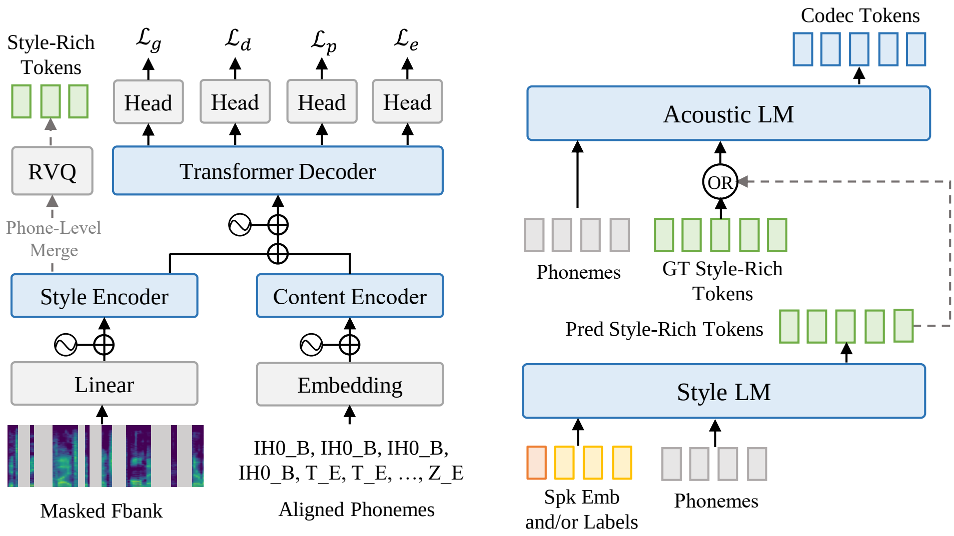
2.2 Speech style representations

Various works attempt to obtain style representations of speech at different granularities with disentanglement or other methods to facilitate voice conversion, controllable TTS, and other applications. NANSY (Choi et al., 2021) deconstructs input speech into multiple information flows explicitly, and then reconstructs speech from these flows, obtaining a model capable of voice conversion, pitch shift, and other applications. SpeechSplit 1 and 2 (Qian et al., 2020; Chan et al., 2022) disentangle speech into content, rhythm, pitch, and timbre using multiple autoencoders in an unsupervised manner. DSVAE (Lian et al., 2022b,a, 2023) presents a self-supervised method to disentangle content information and global speaker information, in an end-to-end manner. Prosody-TTS (Huang et al., 2023) utilizes an MAE to learn a prosody representation disentangled from content and speaker timbre, boosting expressive TTS. NaturalSpeech 3 (Ju et al., 2024) proposes a codec that factorizes speech into individual subspaces representing different attributes like content, prosody, timbre, and acoustic details, facilitating the modeling of intricate speech. In this paper, we adopt a masked autoencoder to extract speech features with rich style information, which are then used as an intermediary to facilitate controllable TTS.

3 Method

3.1 Overview

Our controllable TTS system consists of two major stages with a discrete style-rich token as an intermediate representation. This style-rich representation is from a transformer-based MAE as illustrated in figure 1 (a), which learns to capture style information including speaker timbre, prosody, and acoustic environment in the speech with a mask-reconstruction paradigm. The style-rich tokens of a speech clip can be extracted with the style encoder of the pre-trained MAE followed by a residual vector quantizer (RVQ) trained individually. The two stages of TTS are (1) **style-rich token (ST) generation**, which generates style-rich tokens conditioned on content phonemes and style controlling signals including discrete labels and / or continuous speaker embeddings; and (2) **codec token (CT) generation**, which generates codec tokens conditioned on content phonemes and style-rich tokens, where the style-rich tokens are either extracted from ground truth speech or predicted by



(a) Style MAE (b) Controllable Text-to-Speech Pipeline

Figure 1: Model overview of our controllable TTS system. Figure (a) shows the architecture of the style MAE. Figure (b) illustrates the two-stage controllable TTS pipeline. The gray dashed lines represent paths that occur only during inference.

the former stage. The generated codec tokens are then used to reconstruct the waveform with the codec decoder. Each of the two stages relies on a decoder-only transformer to conduct LM-style generation, as illustrated in figure 1 (b). We provide details of these modules respectively in the following subsections. Details of model configurations are provided in appendix A.

3.2 Style masked autoencoder and feature tokenization

The style masked autoencoder aims to learn to extract style information like speaker timbre, prosody, and acoustic environment by reconstructing mel filterbanks from masked ones and an additional content input with reconstruction and several auxiliary losses. Its architecture is illustrated in figure 1 (a). The two branches of input, which are masked fbanks and a temporal-aligned phoneme sequence where each phoneme is duplicated by its duration, are processed by two encoders separately. Both the style encoder and content encoder are multi-layer transformer encoders. The output of the two encoders together with sinusoidal positional embedding are added and fed to the transformer decoder.

Following Huang et al. (2023), we append four different linear heads at the end of the decoder for output projection used for different optimization objectives. The four objectives are (1) reconstruction loss $\mathcal{L}_r$: mean square loss between the masked fbank patches and the output of the reconstruction head; (2) contrastive loss $\mathcal{L}_c$: InfoNCE loss to maximize the similarity between the head output and the corresponding fbank patch, while minimizing its similarity with non-corresponding fbank patches; (3) pitch classification loss $\mathcal{L}_p$ and (4) energy classification loss $\mathcal{L}_e$, which are cross-entropy losses calculated on log-scale fundamental frequency (f0) and the L2-norm of the amplitude spectrogram from short-time Fourier transform, respectively, both of which are frame-level and binned to 256 scales. The final loss is a linear combination of the four losses:

$$\mathcal{L} = \lambda_r \mathcal{L}_r + \lambda_c \mathcal{L}_c + \lambda_p \mathcal{L}_p + \lambda_e \mathcal{L}_e \quad (1)$$

where $\lambda_r = 10$, and λ_c, λ_p, λ_e are all 1. Intuitively, this design enables the MAE to extract content information from the encoded feature of the aligned phonemes, while extracting style information from the encoded feature of the masked fbank for reconstruction. Once the MAE finishes training, its style encoder should be able to capture various style information from speech.

To reduce the sequence length for language modeling and eliminate redundant information in the style features, we conduct phone-level merge by averaging the frame-level features in the range of each phoneme. After that, we train an RVQ with 3 codebooks independently over the phone-level style features for discretizing the style-rich representation for LM-style modeling. Note that such an architecture and training approach cannot fully prevent content information from leaking into the representations extracted by the style encoder, as it does not include a suitable bottleneck or supervisory signal to achieve this. This is why we refer to it as *style-rich* token rather than *style* token. Nevertheless, this does not hinder the effectiveness of this representation in subsequent TTS applications.

3.3 Two-stage LM-style controllable text-to-speech

We use a decoder-only transformer for autoregressive generation for each of the two stages. Specifically, we adopt the multi-scale transformer as the backbone model (Yang et al.; Huang et al., 2024), which utilizes a stacked global-local transformer architecture to handle multi-codebook token modeling and has exhibited remarkable capabilities in audio synthesis. Details of the model architecture are provided in appendix B. During training, the conditional inputs and target outputs are concatenated into a single sequence and fed to the transformer, with each part having a modality-specific *start* and *end* token at both ends. The LMs model the conditional distribution using next-token prediction with cross-entropy loss calculated on the target output part.

ST Generation In the first stage, we adopt a style LM to generate style-rich tokens from phonemes and control signals. This procedure can be formulated as:

$$\mathrm{P}(\mathbf{s}) = \prod_{t=1}^{T}\prod_{i=1}^{N} \mathrm{P}(s_t^i|\tau, c, \mathbf{s}_{<t}, \mathbf{s}_t^{<i}; \theta_s) \quad (2)$$

where $\mathbf{s}$, τ, c, and θ_s are style-rich tokens, phonemes, control signals, and model parameters, respectively. Here, the control signals can be a speaker embedding and / or discrete control labels. For discrete control labels, we include *age*, *gender*, *pitch mean* for average pitch, *pitch std* for the extent of pitch variation, emotion represented by *arousal*, *dominance*, and *valence*, *SNR* for signal-noise rate, and *C50* for reverberation level. These labels are denoted by extracting attribute values with some tools and binning them to different levels. We can use all these labels to generate speech with a new speaker, or combine part of them like emotion labels with a speaker embedding to adjust these attributes on the basis of a reference speaker. The training data of this stage can be scaled up to large corpora to achieve higher style diversity and control accuracy.

CT Generation In the second stage, we adopt an acoustic LM to generate codec tokens from phonemes and style-rich tokens. No additional control signal is used in this stage, as we assume that the style information is carried by the style-rich tokens. During training, the model takes ground truth style-rich tokens and learns to reconstruct codec tokens of the speech. In inference, the style-rich tokens can be either ground truth ones for speech reconstruction, or predicted ones from the former stage for controllable TTS. This procedure can be formulated as:

$$\mathrm{P}(\mathbf{a}) = \prod_{t=1}^{T}\prod_{i=1}^{N} \mathrm{P}(a_t^i|\tau, \mathbf{s}, \mathbf{a}_{<t}, \mathbf{a}_t^{<i}; \theta_a). \quad (3)$$

where $\mathbf{a}$, τ, $\mathbf{s}$, and θ_a are codec tokens, phonemes, style-rich tokens, and model parameters, respectively. We observe in our experiment that several hundred hours of data are sufficient for the model to learn to reconstruct speech of decent quality from phoneme and style-rich tokens, therefore addressing the scarcity issue of high-quality corpora for controllable TTS.

3.4 Classifier-free guidance

We observe that for attributes with distinct differences among categories (like gender), simply adding the label to the prefix condition sequence leads to pretty good control capability. However, for attributes with fine-grained levels and relatively ambiguous boundaries, this simple approach leaves room for improvement in control accuracy. To enhance the model's control capabilities, we introduce classifier-free guidance (CFG) (Ho and Salimans, 2021), which is initially used in score-based generative models and performs well in aligning conditional input and results. We investigate CFG in the ST generation stage.

Specifically, during the training of the style LM, we randomly replace the controlling labels with a special empty control token with a probability of $p = 0.15$. During inference, for each position i, we apply correction to the logit value of style-rich token s_i with the formula

$$\begin{aligned}
&\log \hat{\mathrm{P}}(s_t^i|\mathbf{s}_{<t}, \mathbf{s}_t^{<i}, \tau, c; \theta_s) \\
&= \gamma \log \mathrm{P}(s_t^i|\mathbf{s}_{<t}, \mathbf{s}_t^{<i}, \tau, c; \theta_s) \quad (4) \\
&+ (1-\gamma) \log \mathrm{P}(s_t^i|\mathbf{s}_{<t}, \mathbf{s}_t^{<i}, \tau, \varnothing; \theta_s)
\end{aligned}$$

where τ, c, and γ represent text (phonemes), control labels, and guidance scale, respectively. The re-calculated logit is then used for calculating the probability for sampling with the softmax function. Appropriate CFG scales improve the style coherence between the generated speech and the fine-grained control labels, boosting the control capability of the model to some extent. Note that we conduct only CFG on discrete control labels but not on speaker embeddings.

4 Experiments

4.1 Dataset and style attributes labeling

We adopt large-scale corpora for training the style MAE, where we combine Gigaspeech-xl (Chen et al., 2021) and Librispeech (Panayotov et al., 2015). We use Gigaspeech-xl solely for training the style LM, and use high-quality LibriTTS (Zen et al., 2019) with a relatively small scale for training the acoustic LM. For evaluation, we randomly pick small sets of samples respectively from LibriTTS (184 samples), GigaSpeech (173 samples), and a dialogue dataset, DailyTalk (Lee et al., 2023) (201 samples), to evaluate the models' performance across different data domains.

To train the style LM, we need to label the different attributes of the data. We utilize multiple annotation tools to extract continuous values or classification probabilities for different speech attributes, and split them into different bins by performing equidistant division within an upper and lower boundary that covers most of the data to obtain the discrete control labels. Details of labeling tools and splitting strategies are provided in appendix C. Besides, considering the correlations between control signals, we discuss methods to determine the range of low-level label intervals from high-level labels to reduce information conflicts in appendix D.

4.2 Metrics

Our evaluation of model performance primarily consists of speech naturalness, content accuracy, speaker similarity, speech reconstruction quality, and control accuracy. We adopt different objective metrics for evaluation. For speech naturalness, we adopt UTMOS (Saeki et al., 2022) to predict the MOS score of each sample and report mean values and 95% confidence intervals for each test set. For content accuracy, we use Whisper large-v3 (Radford et al., 2022) to transcribe the speech and calculate the word error rate (WER) against the ground truth text. For speaker similarity, we compute cosine similarity on speaker embedding extracted by wavlm-base-plus-sv [1]. For reconstruction quality, we calculate MCD between generated and ground truth speech with tools provided in fairseq [2]. For control accuracy, we use the annotation tools to extract attribute labels and compute percentage accuracy with ground truth labels. Considering the challenges of achieving precise control with fine-grained labels, we make some relaxation that results differing from the ground truth attribute label by one bin are also considered correct for *age*, *SNR* and *C50*, and are considered half correct (taken as 0.5 correct samples) for emotion and pitch labels.

We also conduct subjective evaluations and report mean-opinion-scores of speech naturalness (MOS-Q), style alignment with control labels (MOS-A), and timbre similarity with the reference speaker (MOS-S). Details of subjective metrics are provided in appendix E.

4.3 Results and analysis

4.3.1 Reconstruct speech style from style-rich tokens and phonemes

To validate that our style-rich tokens encapsulate rich voice style information, we reconstruct speech from phonemes and ground truth (GT) style-rich tokens, and compare them with original speech, compressed speech from the codec, and zero-shot TTS results. We use YourTTS (Casanova et al., 2022) and XTTS-V2 (Casanova et al., 2024) as representative zero-shot TTS systems for comparison. The results on LibriTTS and Gigaspeech are shown in table 1. For results on both test sets, our model achieves comparable UTMOS to recent zero-shot TTS systems. This demonstrates the reliability of our model in terms of speech naturalness. Besides, our model achieves comparable speaker similarity with zero-shot TTS systems, indicating that the style-rich tokens contain rich speaker information for speech synthesis. Moreover, the reconstruction results have significantly lower MCD than zero-shot TTS, proving that it is closer to the original audio in terms of prosody and other style information like acoustic environment, which further

[1] https://huggingface.co/microsoft/wavlm-base-plus-sv

[2] https://github.com/facebookresearch/fairseq/blob/main/examples/speech_synthesis/docs/ljspeech_example.md#mcdmsd-metric

Table 1: Comparing reconstructed speech from phonemes and ground truth style-rich tokens to original speech, compressed speech and zero-shot TTS results.

	LibriTTS			Gigaspeech		
Method	SIM	MCD	UTMOS	SIM	MCD	UTMOS
GT.	/	/	4.06 ± 0.05	/	/	3.47 ± 0.10
GT. + Codec	0.94	1.98	3.43 ± 0.06	0.91	2.21	2.87 ± 0.09
YourTTS	0.91	6.12	3.61 ± 0.09	0.85	6.72	2.33 ± 0.09
XTTS-V2	0.91	5.96	3.68 ± 0.08	0.87	6.48	3.26 ± 0.10
Acoustic LM + GT Style	0.90	3.19	3.63 ± 0.05	0.86	3.68	3.24 ± 0.08

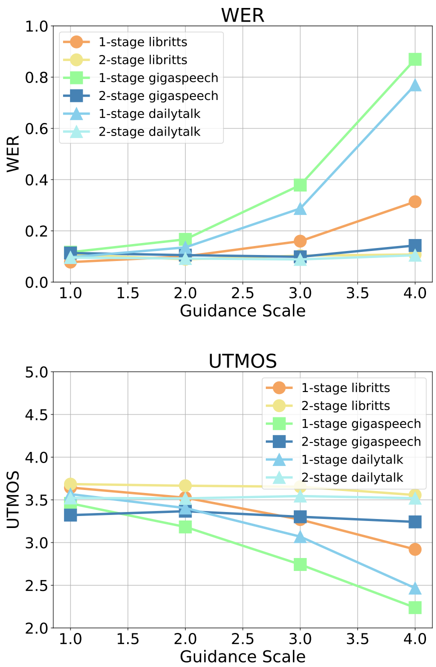


Figure 2: WER and UTMOS on different guidance scales.

validates the effectiveness of our style MAE. We also refer the readers to appendix F for illustration of the reconstructed spectrogram.

4.3.2 Controllable TTS with discrete labels

In this section, we evaluate the performance of our controllable TTS system with solely discrete labels. Considering the differences in control interfaces, target attributes and training data, it is difficult to directly compare our model with previous controllable TTS systems. To validate the effectiveness of our two-stage design, we train a one-stage model as the baseline, which generates codec tokens from phonemes and control labels directly. We use LibriTTS to train the one-stage model, which is the same as training the acoustic LM. Due to the sheer magnitude of their quantity, traversing all possible attribute combinations is not feasible. Furthermore, the correlation among attributes may render certain combinations of labels impossible or difficult to achieve. Therefore, we use label combinations extracted from ground truth speech for control and evaluation and further modify specific attributes for case studies.

We first consider the content accuracy and naturalness of the TTS systems. We illustrate the WER and UTMOS values of the two models under different CFG scales in figure 2. It can be seen that for the one-stage model trained on LibriTTS, as the CFG scale increases, the word error rate rises and UTMOS declines, especially on out-of-domain test sets of Gigaspeech and DailyTalk, manifesting significant degradation in content accuracy and naturalness. This indicates the instability of the one-stage model trained on small, high-quality datasets when subjected to an increased CFG scale, making it difficult to balance control capabilities with speech quality. On the other hand, the two-stage model with the first stage trained on large corpora exhibits good and stable content accuracy and naturalness with growing CFG scales. This proves that the first stage trained on extensive data helps in enhancing the content robustness of controllable TTS, without affecting speech quality by error propagation.

In figure 3, we illustrate the control accuracies of the two-stage model under different CFG scales. We can see that the effect of CFG varies for different attributes. For gender attributes with fewer categories and significant differentiation, the presence or absence of CFG shows no clear impact and the model achieves good control performance in both cases. However, for fine-grained attributes like *arousal* and *pitch mean*, appropriate CFG scales can benefit control accuracy, especially on LibriTTS and DailyTalk test sets. This indicates that CFG helps in the precise control of fine-grained attributes. Meanwhile, we find that larger CFG scales are not always beneficial. For some attributes, control accuracy initially increases before subsequently

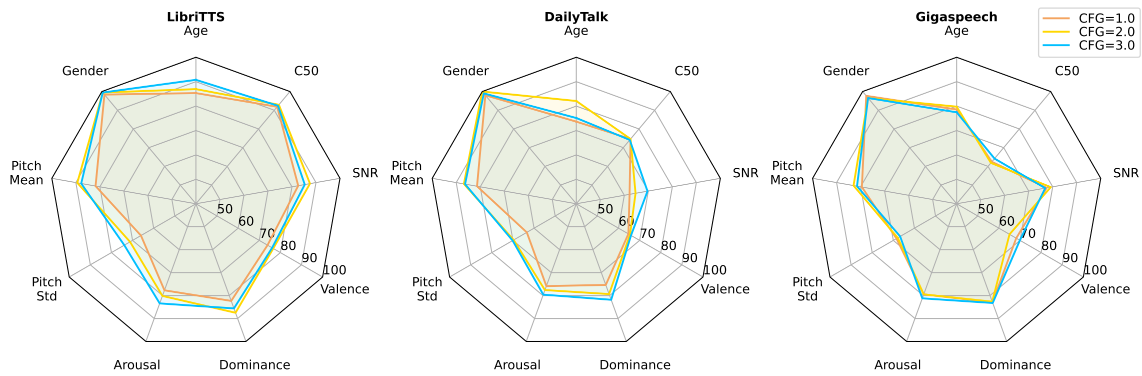


Figure 3: Control accuracy of the two-stage controllable TTS with discrete labels under different CFG scales. The coordinate range is also set to 40-100.

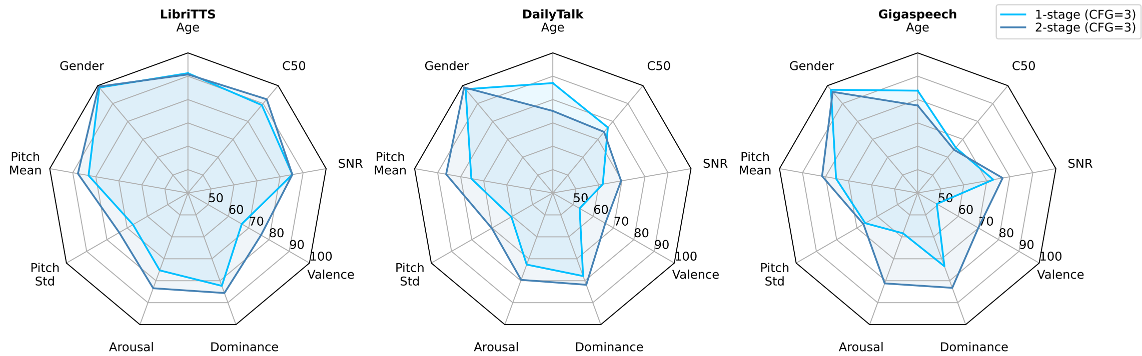


Figure 4: Control accuracy of the one-stage and two-stage controllable TTS with discrete labels under a CFG scale of 3.0. The coordinate range is set to 40-100 for the more apparent differences.

declining as the scale rises. We speculate that this may be due to larger scale values causing distortion in the generated speech, similar to the phenomenon observed with CFG in score-based models. The full results of these two models under different CFG scales are provided in appendix G.

We further evaluate the control ability of the models. In figure 4, we compare the control accuracies of the one-stage and the two-stage model under a CFG scale of 3.0. It can be seen that the one-stage model has some advantages in *age* control, while the two-stage model achieves comparable or superior control over other attributes. The two-stage model shows significant advantages in emotion control and average pitch, and it also achieves better accuracy over pitch variation and *SNR* on part of the test sets. This indicates that compared to the one-stage model trained on high-quality corpora with limited scale, the two-stage model with the first stage trained with extensive data boosts modeling diverse pitch and acoustic conditions. We refer the readers to appendix F for spectrogram samples that intuitively demonstrate the model's control capabilities.

4.3.3 Controlling pitch and emotion with a reference speaker

In this section, we present the results that alternate the timbre-related labels including *age* and *gender* with speaker embedding from WeSpeaker (Wang et al., 2023b) to achieve control over emotion attributes with a specified reference speaker, and investigate the emotion control capability of the model. The pitch and acoustic condition labels are kept in the condition sequence. We present results on Gigaspeech and DailyTalk in table 2. It can be seen that our model achieves decent speaker similarity on both test sets as well as comparable control accuracy to the discrete-label-only paradigm over emotion. This indicates the effectiveness of our model in controlling emotion for a specified speaker. Moreover, compared to fully-discrete-label controlling, the one-stage model shows better content robustness with growing CFG scale in this setting, and the one-stage and two-stage models exhibit comparable performance in content accuracy and speaker similarity. Despite this, the two-stage model retains advantages in control over the emotional attributes, demonstrating that the ST generation model trained on an extensive dataset remains

Table 2: Results of controllable TTS combining speaker embedding, pitch and emotion labels.

Test set	Model	CFG Scale	WER(%)	SIM	Aro.	Dom.	Val.	UTMOS
Gigaspeech	1-stage	1.0	0.13	0.85	69.1	74.9	63.0	3.33 ± 0.08
		2.0	0.12	0.85	73.1	77.7	67.1	3.30 ± 0.08
		3.0	0.14	0.86	70.5	75.7	62.1	3.27 ± 0.07
	2-stage	1.0	0.12	0.86	76.9	76.3	68.5	3.24 ± 0.09
		2.0	0.14	0.85	78.0	78.6	68.5	3.26 ± 0.09
		3.0	0.14	0.86	76.0	80.9	65.6	3.24 ± 0.09
DailyTalk	1-stage	1.0	0.14	0.82	65.7	71.6	58.5	3.28 ± 0.07
		2.0	0.13	0.82	66.7	71.6	59.5	3.24 ± 0.07
		3.0	0.15	0.82	68.9	75.1	59.0	3.18 ± 0.07
	2-stage	1.0	0.10	0.80	73.9	78.6	62.7	3.50 ± 0.07
		2.0	0.10	0.80	76.1	81.6	64.4	3.53 ± 0.07
		3.0	0.09	0.80	79.6	83.3	63.7	3.51 ± 0.07

advantageous in modeling pitch-related stylistic information in this setting.

4.3.4 Subjective evaluation on model performance

Table 3: Subjective evaluation results.

Model	CFG Scale	MOS-Q	MOS-A	MOS-S
Control with discrete labels				
1-stage	1.0	4.11 ± 0.11	3.99 ± 0.13	/
	2.0	3.81 ± 0.12	3.98 ± 0.11	/
	3.0	2.89 ± 0.14	3.45 ± 0.13	/
2-stage	1.0	4.14 ± 0.13	3.93 ± 0.13	/
	2.0	4.01 ± 0.11	4.20 ± 0.14	/
	3.0	4.18 ± 0.12	4.20 ± 0.11	/
Control with speaker embeddings and emotion labels				
1-stage	1.0	3.96 ± 0.12	3.61 ± 0.13	3.89 ± 0.11
	2.0	3.90 ± 0.11	3.90 ± 0.14	3.58 ± 0.14
	3.0	3.70 ± 0.12	3.86 ± 0.12	3.40 ± 0.13
2-stage	1.0	3.97 ± 0.12	4.06 ± 0.13	3.56 ± 0.12
	2.0	4.13 ± 0.11	4.23 ± 0.12	3.68 ± 0.12
	3.0	3.91 ± 0.11	4.28 ± 0.10	3.52 ± 0.13

Table 3 presents the results of our subjective evaluations. As shown, the two-stage model demonstrates comparable or superior MOS-A to the one-stage model, indicating its superior control capabilities. Additionally, an appropriate CFG scale leads to better control performance. Meanwhile, for the one-stage model trained with a small dataset, increasing the CFG scale while using only the labels as the control signal leads to a decrease in MOS-Q. These results align with the conclusions reflected by the objective metrics.

5 Conclusion

In this paper, we propose an LM-based fine-grained controllable TTS system. We adopt a two-stage generation pipeline, with an autoregressive transformer as the backbone for each stage. We design a masked autoencoder for extracting features with rich style information from the speech and use the discretized feature as the intermediate output of the TTS pipeline. By selectively combining discrete control labels with speaker embeddings, our model supports both generating new speaker timbre while controlling other attributes, and controlling emotion for a specified speaker. Experiments indicate the effectiveness of our model.

In the future, we may explore more diverse control signals and employ techniques such as prompt engineering to integrate large language models with controllable TTS, enabling support for both natural language prompts and fine-grained control signals.

6 Limitations

Despite that our approach achieves fine-grained control over multiple style attributes, our method and evaluation protocols still suffer from several limitations: 1) Due to the performance limitations of labeling tools, there may be errors in the attribute annotations of the training data, which could lead to a decline in the model's control capabilities. 2) Evaluation with label combinations from real data may present issues of uneven distribution, particularly for attributes with significant distribution bias, such as SNR and C50. Therefore, the evaluation may not fully accurately reflect the model's control capabilities. 3) Due to their small proportion in the training data, some marginal labels and their combinations may lead to degraded generated audio and diminished control performance. We will explore solutions to these issues in future work.

7 Potential Risks

Improper use of this model may lead to the creation of fake content, such as generating statements that a specific speaker has never made. It may also cause copyright issues. We will add some constraints to guarantee people who use our code or pre-trained model will not use the model in illegal cases.

References

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, et al. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR.

Chak Ho Chan, Kaizhi Qian, Yang Zhang, and Mark Hasegawa-Johnson. 2022. Speechsplit2. 0: Unsupervised speech disentanglement for voice conversion without tuning autoencoder bottlenecks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6332–6336. IEEE.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Hyeong-Seok Choi, Juheon Lee, Wansoo Kim, Jie Lee, Hoon Heo, and Kyogu Lee. 2021. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. *Advances in Neural Information Processing Systems*, 34:16251–16265.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jonathan Ho and Tim Salimans. 2021. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Rongjie Huang, Chunlei Zhang, Yi Ren, Zhou Zhao, and Dong Yu. 2023. Prosody-tts: Improving prosody with masked autoencoder and conditional diffusion model for expressive text-to-speech. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8018–8034.

Rongjie Huang, Chunlei Zhang, Yongqi Wang, Dongchao Yang, Jinchuan Tian, Zhenhui Ye, Luping Liu, Zehan Wang, Ziyue Jiang, Xuankai Chang, Jiatong Shi, Chao Weng, Zhou Zhao, and Dong Yu. 2024. Make-a-voice: Revisiting voice large language models as scalable multilingual and multitask learners. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10929–10942.

Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. 2024. Textrolspeech: A text style control speech corpus with codec language text-to-speech models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10301–10305. IEEE.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yichong Leng, Zhifang Guo, Kai Shen, Xu Tan, Zeqian Ju, Yanqing Liu, Yufei Liu, Dongchao Yang, Leying Zhang, Kaitao Song, et al. 2023. Prompttts 2: Describing and generating voices with text prompt. *arXiv preprint arXiv:2309.02285*.

Jiachen Lian, Chunlei Zhang, Gopala K. Anumanchipalli, and Dong Yu. 2023. Unsupervised tts acoustic modeling for tts with conditional disentangled sequential vae. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2548–2557.

Jiachen Lian, Chunlei Zhang, Gopala Krishna Anumanchipalli, and Dong Yu. 2022a. Towards improved zero-shot voice conversion with conditional dsvae. In *ISCA Interspeech*.

Jiachen Lian, Chunlei Zhang, and Dong Yu. 2022b. Robust disentangled variational speech representation learning for zero-shot voice conversion. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6572–6576. IEEE.

Dan Lyth and Simon King. 2024. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *arXiv preprint arXiv:2402.01912*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox. 2020. Unsupervised speech decomposition via triple information bottleneck. In *International Conference on Machine Learning*, pages 7836–7846. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. arxiv 2022. *arXiv preprint arXiv:2212.04356*, 10.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. 2023. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. 2023b. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Wikipedia. 2024. Pearson correlation coefficient — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Pearson%20correlation%20coefficient&oldid=1246912561. [Online; accessed 01-October-2024].

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. 2024. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Haohan Guo, Xuankai Chang, Jiatong Shi, Jiang Bian, Zhou Zhao, et al. Uniaudio: Towards universal audio generation with large language models. In *Forty-first International Conference on Machine Learning*.

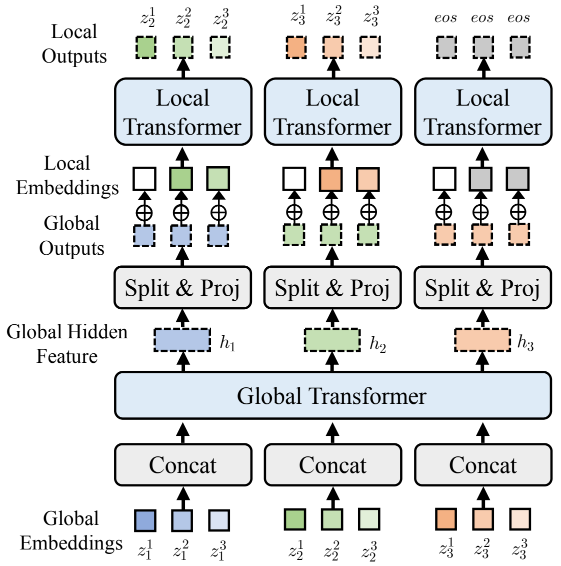


Figure 5: Illustration of the multi-scale transformer backbone.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Yixuan Zhou, Xiaoyu Qin, Zeyu Jin, Shuoyi Zhou, Shun Lei, Songtao Zhou, Zhiyong Wu, and Jia Jia. 2024. Voxinstruct: Expressive human instruction-to-speech generation with unified multilingual codec language modelling. *arXiv preprint arXiv:2408.15676*.

Table 4: Hyper-parameters of different modules of our approach.

Model	Hyperparameter	
Style MAE	Encoder Layers	12
	Decoder Layers	2
	Hidden Dimension	768
	Mask Probability	0.75
	Fbank Channels	128
Style LM & Acoustic LM	Global Layers	20
	Local Layers	6
	Hidden Dim	1,152
	Global Attention Heads	16
	Local Attention Heads	8
	FFN Dim	4,608

A Implementation details

In table 4, we illustrate the model hyper-parameters of the style MAE and two language models in our approach. For codec, we train a EnCodec (Défossez et al., 2022) model for 16k audio, with 8 quantization levels, a codebook size of 1024, and a downsampling rate of 320. We use the first 3

Table 5: Extracting tools and binning strategies for different attributes.

Attribute	Extracting Tool	Lower Bound	Upper Bound	Bin Number
Gender	w2v2-age-gender	0.0	1.0	4
Age	w2v2-age-gender	0	100	10
Arousal, Dominance, Valence	w2v2-emotion	0.2	0.8	7
Pitch Mean	DataSpeech	45.0	320.0	10
Pitch Std	DataSpeech	0.0	132.0	10
SNR	DataSpeech	-9.16	77.13	10
C50	DataSpeech	0.0	25.0	10

quantization levels only. We also use 3 RVQ layers for style-rich tokens.

B Multi-scale transformer architecture

The hierarchical structure of the multi-scale transformer is illustrated in figure 5. This structure is formed by a global and a local transformer, both of which are decoder-only transformers. For a temporal position t, embeddings $z_t^{1:n_q}$ of style-rich or acoustic tokens from different codebooks are concatenated and fed to the global transformer for inter-frame correlation modeling. The output hidden feature h_t is generated autoregressively conditioned on $h_{1:t-1}$. This hidden feature is then split according to the original shape of the embeddings, projected by a linear layer, and added to the input embeddings of the local transformer as a frame-level context. The local transformer predicts style-rich or acoustic tokens of different codebooks inside a frame autoregressively. For other modalities, each item is repeated n_q times to fit this modeling mechanism, with n_q being the number of codebooks.

C Style attribute labeling

In this section, we provide details of how we obtain the labels of different attributes. The extracting tools and binning strategies are summarized in table 5. For age and gender, we use a finetuned wav2vec2 model [3] to extract gender classification probability and estimated age between 0-100. We then split age into 4 categories: *male*, *neutral-masculine*, *neutral-feminine*, *female*, with the criteria being the probability of *male*, and thresholds of 0.65, 0.5 and 0.35.

For emotion labels, we adopt another finetuned wav2vec2 model [4] to extract the predicted logits of arousal, dominance, and valence. The range of the logits is 0-1, yet most audio falls between 0.2 and 0.8. Therefore, we divide the interval from 0.2 to 0.8 into seven labels with a distance of 0.1.

For pitch and acoustic conditions, we utilize DataSpeech (Lyth and King, 2024) to extract the mean value and standard variation of pitch, as well as SNR and C50. The ranges between the upper and lower bounds of each attribute are divided into 10 equidistant intervals, with the boundaries listed in the table.

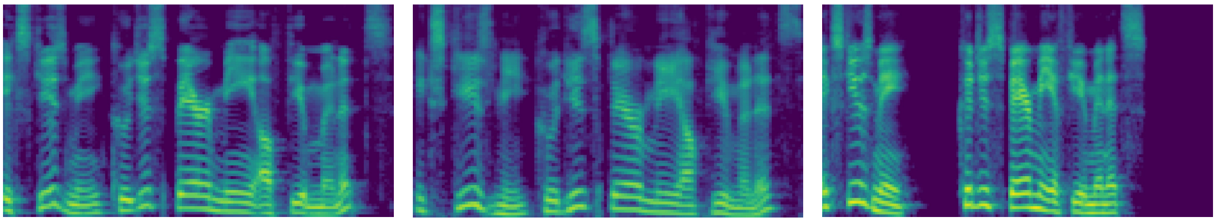
D Correlation among control attributes

In fact, the information contained among different attributes may overlap, manifesting as correlations between labels. Certain high-level attributes can be reflected in lower-level acoustic properties. For example, attributes related to speaker timbre, such as age and gender, are closely linked to average pitch, while emotion is closely related to pitch variation. In table 6, we present the Pearson correlation coefficients (Wikipedia, 2024) between high-level attributes and pitch attributes calculated on LibriTTS. It can be seen that age is correlated with average pitch to some degree, while gender, arousal, and dominance show significant correlations with both the mean and variation of pitch, indicating the presence of overlapping information. Additionally, the limited performance of the annotation tools may also lead to significant correlation among different emotional dimensions. Theoretically, the three dimensions of arousal, dominance, and valence are orthogonal. However, as shown in figure 7, the distributions of arousal and dominance extracted by the model exhibit a strong linear correlation.

Due to the correlation among different attributes, using control signals that contain conflict information may lead to sub-optimal speech quality and control capability. We showcase examples on our demo page where conflicting control signals lead to degraded control performance. To achieve better control accuracy and content quality, we can

[3] https://github.com/audeering/w2v2-age-gender-how-to

[4] https://github.com/audeering/w2v2-how-to



(a) Ground Truth (b) Acoustic LM + GT Style (c) XTTS-V2

Figure 6: Spectrogram from original speech, reconstructed speech with ground truth style-rich tokens and zero-shot TTS result.

Table 6: Pearson correlation coefficients between high-level and low-level attributes.

Low-Level \ High-Level	Age	Gender	Arousal	Dominance	Valence
Pitch Mean	-0.15	-0.74	0.38	0.29	0.06
Pitch Std	-0.01	0.37	0.39	0.33	0.06

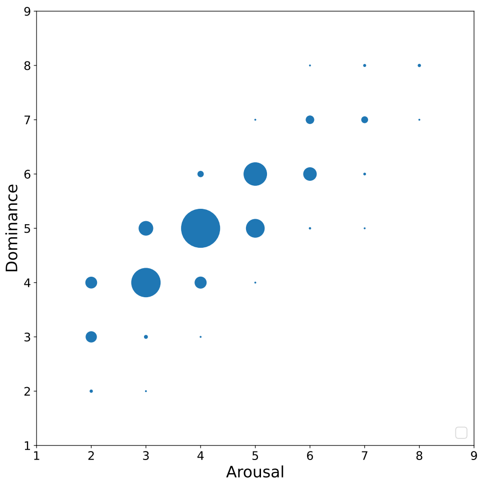


Figure 7: Illustration of the data distribution for arousal and dominance.

restrict the ranges of low-level attributes with desired high-level attribute labels, thereby avoiding information conflicts. A straightforward solution is a statistical approach, where we can calculate the conditional distributions of *pitch mean* and *pitch std* given other labels on the training dataset, and sample labels from the distribution. Another solution is a learning-based method, where we can train label predictors for estimating low-level attributes from the given high-level labels. We train two 3-layer MLPs with a hidden dimension of 160 to predict *pitch mean* and *pitch std* from *age*, *gender*, *arousal*, *dominance* and *valence*. We find that the accuracy of predicting *pitch mean* and *pitch std* can reach around 40%, while the soft accuracy—considering a label error of no more than 1 as correct—exceeds 80%. This demonstrates the effectiveness of these predictive models. Once these models finish training, the output probabilities can be used to sample pitch labels.

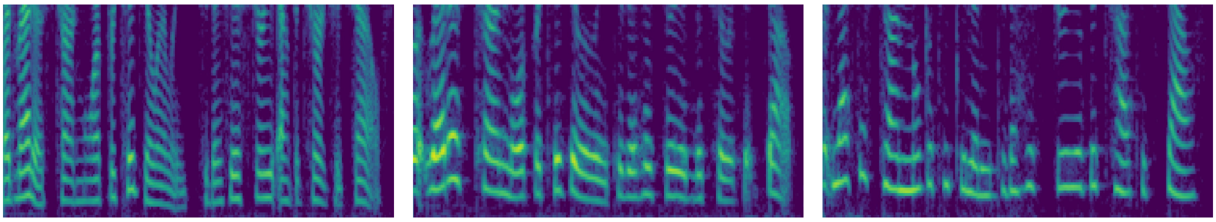
E Subjective Evaluation

We invite 10 individuals with experience in TTS research as participants for our subjective evaluation. For each experiment setting, we select 16 samples for each model for evaluation. The participants rate scores on 1-5 Likert scales, and report mean scores with 95% confidence intervals. For MOS-A, considering that the original VAD labels are difficult to understand, we converted the VAD label combinations into emotional intensity levels (such as *flat*, *neutral*, or *highly expressive*) or typical emotional categories (such as *happy*, *angry*, or *sad*) corresponding to those combinations. The participants are paid $8 hourly.

F Sample illustrations of results

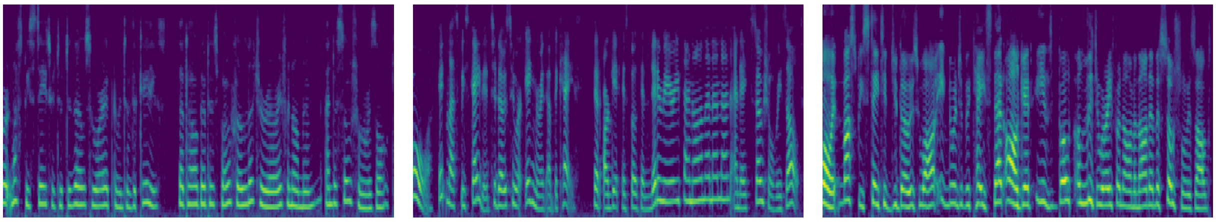
For experiment results in section 4.3.1, figure 6 illustrates the spectrogram of some sample results on DailyTalk. It can be observed that despite some over-smoothing in certain details, the acoustic LM is able to leverage the style information contained in the style-rich tokens to achieve accurate reconstruction on out-of-domain samples, indicating the effectiveness of our style representation. In contrast, zero-shot TTS that only leverages speaker information cannot achieve prosody reconstruction.

To illustrate the control capabilities of the model, we take *pitch mean* and emotion labels as examples,



(a) pitch mean=3 (b) pitch mean=5 (c) pitch mean=7

Figure 8: Spectrograms obtained using pitch labels of different levels in two-stage controllable TTS.



(a) a=2 d=3 v=2 (depressed) (b) a=5 d=6 v=6 (joyful) (c) a=6 d=7 v=2 (angry)

Figure 9: Spectrograms obtained using different compositions of emotion labels in two-stage controllable TTS.

and plot the spectrograms to illustrate the effects of modifying specific attributes of the given samples. Figure 8 showcases the results using different average pitch labels while keeping the content and other attributes constant. We only display the frequency range of 0-2kHz for clearer visualization. It can be seen that when we raise the value of the *pitch mean* label, the fundamental frequency levels up, and the distance between formants increases, indicating that the speaker timbre grows shriller, proving the effectiveness of our model on controlling average pitch. In figure 9, we use three different groups of emotion labels for one test sample. The spectrogram shows that labels corresponding to elevated emotion lead to more pronounced pitch variation compared to those of subdued emotion. We refer the reader to our demo page for more samples.

G Supplementary experiment results

In table 7, we provide the full results of the one-stage and two-stage models with discrete labels under different CFG scales, corresponding to figure 3 and figure 4 in section 4.3.2. This table provides a more accurate and comprehensive comparison of the performance between the one-stage and two-stage models, as well as the impact of CFG scales on both of them. It can be seen that CFG is effective in boosting control performance for both the one-stage and two-stage models. Moreover, the results demonstrate that the two-stage model outperforms the one-stage model in attributes such as pitch mean and arousal across a wide range of settings, further supporting the conclusions drawn in section 4.3.2.

Table 7: Control accuracy of controllable TTS with discrete labels.

Test set	Model	CFG Scale	WER	Age	Gen.	P.M.	P.S.	Aro.	Dom.	Val.	SNR	C50
LibriTTS	1-stage	1.0	0.08	80.4	95.7	74.2	52.7	73.9	77.4	73.6	78.3	87.5
		2.0	0.10	85.9	99.5	82.9	65.5	77.4	88.0	69.8	82.6	90.8
		3.0	0.16	91.3	98.9	83.2	67.1	75.3	82.3	66.6	85.3	89.1
	2-stage	1.0	0.11	85.3	98.4	81.8	66.0	77.7	82.3	73.9	82.6	91.8
		2.0	0.09	87.0	99.5	89.4	71.2	80.2	87.5	76.6	87.5	92.9
		3.0	0.10	90.8	99.5	87.8	73.9	83.4	85.6	76.1	85.3	92.4
Gigaspeech	1-stage	1.0	0.12	70.5	96.0	70.5	57.5	73.4	74.9	62.4	63.0	61.3
		2.0	0.17	82.7	98.3	77.2	63.9	73.7	75.4	59.2	72.3	64.7
		3.0	0.38	83.8	97.7	75.4	65.9	58.4	73.4	49.4	72.8	65.3
	2-stage	1.0	0.11	78.6	97.7	79.5	68.5	79.5	82.9	68.2	77.5	62.4
		2.0	0.11	79.8	96.5	82.9	67.9	79.5	82.7	65.0	79.2	61.8
		3.0	0.10	77.5	96.5	81.5	66.8	81.2	83.2	69.9	76.9	64.2
DailyTalk	1-stage	1.0	0.10	75.6	94.0	70.9	60.4	66.7	71.4	61.4	76.1	73.1
		2.0	0.14	83.1	99.0	75.6	61.4	74.6	78.4	59.5	70.6	72.6
		3.0	0.29	87.1	98.0	75.4	60.4	72.6	77.9	53.2	61.7	76.6
	2-stage	1.0	0.10	73.6	98.0	81.3	63.4	75.9	75.4	64.7	62.2	74.6
		2.0	0.09	82.1	100.0	86.8	69.7	77.6	79.4	65.4	64.7	74.6
		3.0	0.09	75.1	99.0	86.3	70.1	79.6	81.8	65.9	69.7	74.1