Prompt-Singer: Controllable Singing-Voice-Synthesis with Natural Language Prompt

Anonymous ACL submission

Abstract

Recent singing-voice-synthesis (SVS) methods have achieved remarkable audio quality and naturalness, yet they lack the capability to control the style attributes of the synthesized singing explicitly. We propose Prompt-Singer, the first SVS method that enables attribute controlling on singer gender, vocal range and volume with natural language. We adopt a model architecture based on a decoder-only transformer with a multi-scale hierarchy, and design a range-melody decoupled pitch representation that enables text-conditioned vocal range control while keeping melodic accuracy. Furthermore, we explore various experiment settings, including different types of text representations, text encoder fine-tuning, and introducing speech data to alleviate data scarcity, aiming to facilitate further research. Experiments show that our model achieves favorable controlling ability and audio quality. Audio samples are available at http://prompt-singer.github.io.

1 Introduction

Singing-voice-synthesis (SVS) systems (Chen et al., 2020; Liu et al., 2022; Zhang et al., 2022b,c, 2023b), which aim to generate high-fidelity singing voices given lyrics and pitch notes, have made significant advancements in improving audio quality and naturalness in recent years, facilitating music composition and development of entertainment industries. However, it hasn't been fully studied to control the style attributes of synthesized singing, such as speaker timbre, vocal range and energy. Despite that some works use fixed speaker IDs (Huang et al., 2021; Zhang et al., 2022c) or reference speech/singing segments (Shen et al., 2023) to provide information on singer identity or other style attributes, these mechanisms are not user-friendly and lack the ability to control specific acoustic attributes explicitly.

An ideal approach to controlling the style of generated singing voices is to use natural language instructions as style prompts, as it can not only achieve precise control over specific attributes with certain descriptions, but also simplify user interaction, which may bring convenience to non-professional users such as musicians and video creators. However, applying natural language style prompts in singing-voice-synthesis faces several challenges:

- **Decoupling Melody and Vocal Range.** In real-life situations, different speakers (e.g. an elderly man and a little girl) may sing the same song within different vocal ranges. However, pitch annotations in SVS data are each tied to a specific singer in a certain vocal range. This coupling nature makes it challenging to generate singing voices with consistent vocal range and timbre to the prompt together with an accurate melody aligned with given pitch notes.

- **Textual Representation.** Despite that some works have explored connecting text representations with music, speech and general audio concepts (Elizalde et al., 2023a,b; Wu et al., 2023), there is no text representation tailored for singing style descriptions, and the optimal choice of prompt representation for this task remains unknown.

- **Data Scarcity.** Due to the requirement of fine-grained annotations, existing SVS datasets (Liu et al., 2022; Wang et al., 2022; Huang et al., 2021; Zhang et al., 2022a) are small in scale, typically consisting of only a few hours or tens of hours of singing data. This not only causes limited data diversity but also poses more challenges to learning the correlation between natural language descriptions and data distribution.

In this paper, we propose Prompt-Singer, the first controllable SVS model with natural language prompts to control the singer gender, vocal range

and volume. Considering the outstanding performance of recent spoken LLMs (Borsos et al., 2023; Wang et al., 2023; Yang et al., 2023b) in terms of generation and in-context learning capabilities, we adopt a decoder-only transformer with a multi-scale hierarchy for conditional generation of discrete codec units of the singing, together with a unit vocoder for waveform reconstruction. To address the challenges mentioned above, we 1) design a decoupled pitch representation with a vocal range factor and a speaker-independent melody sequence, enabling voice range controlling while maintaining melodic accuracy; 2) investigate various text encoders for prompt encoding, as well as fine-tuning the encoders to seek the optimal textual representation for this task; 3) introduce speech data to alleviate data scarcity, and evaluate the model performance under different levels of low-resource singing data combined with speech data. Experiments show that our method achieves favorable style controlling accuracy on the three attributes, while keeping good audio quality and melodic accuracy. Our contributions are summarized as follows:

- We propose the first controllable SVS model with natural language prompts to control the singer gender, vocal range, and volume of the generated singing voice.
- We design a pitch representation for SVS that decouples voice range and melody, which enables prompt-conditioned voice range manipulation while keeping melodic accuracy.
- We investigate different text representations and fine-tune the text encoders to seek optimal text representation for the prompt in this task.
- We alleviate data scarcity by introducing speech data, which boosts prompt-SVS performances in low-resource scenarios.

2 Related Works

2.1 Singing Voice Synthesis

Singing-voice-synthesis aims to generate human-like singing voices from lyrics and pitch notes, and recent deep-learning-based models have achieved remarkable progress in synthesized voice quality. Several works (Chen et al., 2020; Zhang et al., 2022c, 2023b) adopt generative adversarial networks for high-fidelity SVS. Diffsinger (Liu et al., 2022) adopts a shallow diffusion mechanism to enhance the quality of the generated mel-spectrogram. VISinger (Zhang et al., 2022b) proposes an end-to-end architecture based on a variational autoencoder. However, it has not been fully studied to control the style of generated singing. Previous multi-singer systems (Huang et al., 2021; Zhang et al., 2022c) use a fixed group of IDs to indicate singer identities. NaturalSpeech 2 (Shen et al., 2023) uses a reference singing or speech clip to provide holistic style information. Currently, there is a lack of fine-grained controllable methods for SVS.

2.2 Instruct-guided Voice Generation

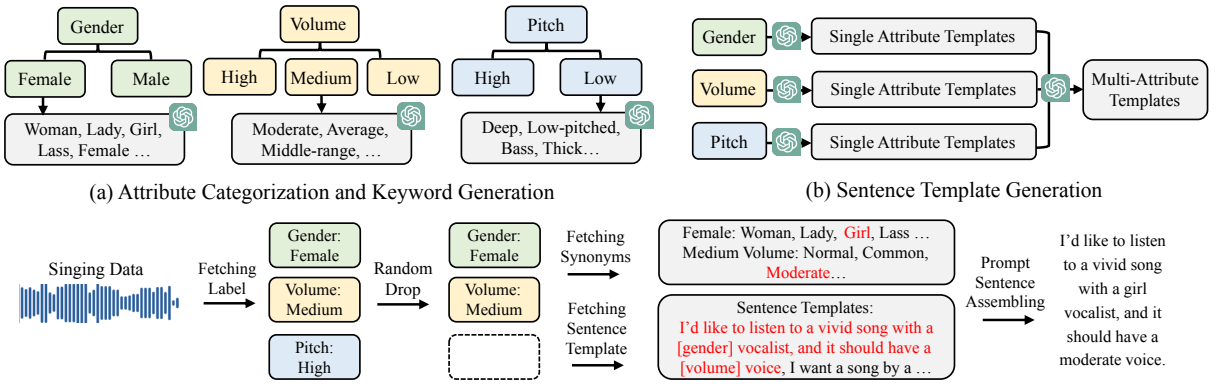
Inspired by the success in text, image and audio generation guided with natural language instructions (Brown et al., 2020; Ramesh et al., 2021; Kreuk et al., 2022), some recent works have explored using text prompts to govern the stylistic attributes in voice synthesis. PromptTTS (Guo et al., 2023) incorporates style features from a fine-tuned BERT into a TTS backbone with attention. InstructTTS (Yang et al., 2023a) 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. PromptVC (Yao et al., 2023) and PromptSpeaker (Zhang et al., 2023a) investigate text-prompted voice conversion and speaker-embedding generation separately. However, due to the data scarcity and the demand for precise pitch controlling, research on natural-language-instructed SVS is currently lacking.

3 Prompt Generation and Fetching

Our goal is to control the singer gender, vocal range and volume in singing-voice-synthesis with natural language prompts. Since there is no available dataset for this task, we utilize normal SVS datasets and design a method for generating a prompt sentence for each data item. We introduce this process in this section.

Considering the high cost of manual annotation, we utilize a large language model (GPT 3.5 Turbo) to generate prompt sentences. The prompt generation mainly consists of 3 stages: 1) attribute categorization; 2) keyword and sentence template generation and 3) prompt sentence assembling.

Figure 1(a) and (b) demonstrate the process of the first two stages. Initially, we categorize the audio based on different attributes. The two gender categories, male and female, are pre-annotated



(a) Attribute Categorization and Keyword Generation

(b) Sentence Template Generation

(c) Prompt Sentence Assembling in Training

Figure 1: The pipeline of generating and fetching prompt sentence for training data.

in the datasets. For volume, we build three categories of "low", "medium", and "high", indicating the amplitude root mean square (RMS) ranges of $[0.02, 0.04]$, $[0.07, 0.10]$ and $[0.16, 0.20]$, respectively. Additionally, we can rescale audio into different ranges dynamically during training. For vocal range, we set two categories of "high" and "low", and use the average f0 of the voiced part as the criterion for classification, with the threshold being 125 Hz for male singers and 305 Hz for female singers.

After categorization, we use LLM to generate a set of 4-7 synonyms for each category as the keywords. We further utilize LLM to generate prompt sentence templates for each single attribute, where each template contains a placeholder to be replaced with the keywords (such as "Generate a song by a [gender] singer"). We also generate a small number of prompt sentences targeting specific categories (such as "Could you synthesize a song that's as powerful as a thunderstorm?" for large volume). We obtain approximately 50 sentence templates for each attribute after manual selection. These single-attribute templates can be further combined to create multi-attribute templates by prompting LLM. We provide sample sentence templates and keywords in Appendix A.

The prompt sentence assembling stage takes place dynamically during training. Figure 1(c) illustrates the pipeline of fetching a prompt sentence. We first obtain the pre-annotated labels for the data item, and in order to make the model adaptable to prompts with varying numbers of attributes, one or two labels are randomly dropped with probabilities p_1 and p_2. We then randomly fetch a keyword and a sentence template from the pre-generated sets, and replace the placeholder with the keyword to get the final prompt sentence. Note that we do not control vocal range independently in the absence of gender, as its boundary is different for male and female. We use pre-generated specific prompts for each sample in the evaluation for fair comparison.

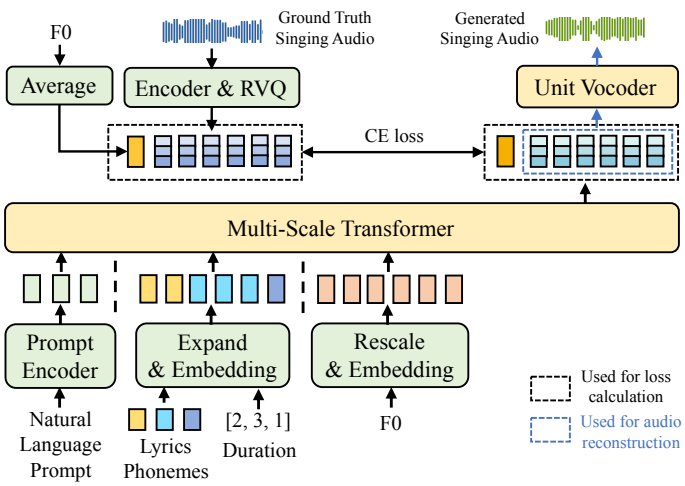
4 Prompt-Singer

In this section, we introduce the model design of Prompt-Singer. The overall architecture of our model is illustrated in Figure 2(a). It is primarily composed of two sub-modules: 1) the multi-scale transformer, which generates discrete acoustic units conditioned on inputs of natural language prompt, lyrics with duration, and pitch information; and 2) the unit vocoder, which maps the generated acoustic units to an audio waveform.

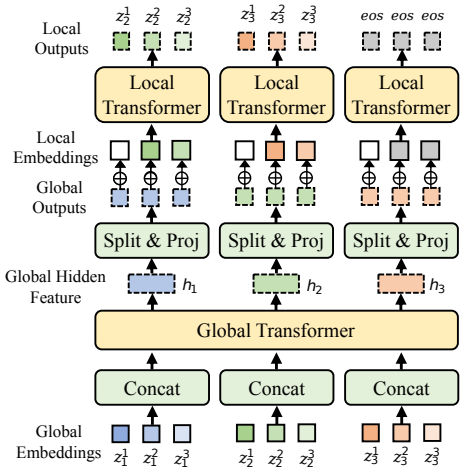
In the following subsections, we introduce the input and output representations of the model in Section 4.1 to 4.3, model architecture in detail in Section 4.5 and 4.6, together with our method for data scarcity alleviation in Section 4.4.

4.1 Voice Representation

The acoustic units used as the prediction targets of the transformer are generated by SoundStream(Zeghidour et al., 2021), a neural codec with an encoder-decoder architecture and a residual vector quantizer (RVQ). Such a codec model can produce discrete compressed representations of audio by employing a convolutional encoder followed by the RVQ, and these representations can be used to reconstruct waveforms with the decoder. An acoustic unit sequence can be represented as $\mathbf{a} = [a_1^1, a_1^2, ..., a_1^C, a_2^1, ..., a_T^C], a_i^j \in \{0, 1, ..., K_a - 1\}, \forall 1 \leq i \leq T, 1 \leq j \leq C$, with



(a) Model Architecture of Prompt-Singer



(b) Multi-Scale Transformer

Figure 2: Model architecture of Prompt-Singer and the multi-scale transformer.

T, C, K_a being number of frames, number of residual codebooks and codebook size.

4.2 Textual Representation

The textual input for our model comprises two components: 1) lyrics, which correspond to the content of the generated song, and 2) natural language prompt, which controls the style of the singing. We introduce their representations in this subsection.

For lyrics, we first phonemize the text and obtain corresponding phoneme-level duration in seconds from dataset annotations or a forced-alignment tool (McAuliffe et al., 2017). We then convert the duration to frame level based on a preset frame rate, and regulate the length of the phoneme sequence with this duration by duplicating phonemes. We set the frame rate of phonemes to be the same as acoustic units, making it easier for the model to learn the length alignment. The regulated phoneme sequence is then embedded by a look-up table (LUT) and fed to the transformer.

For the natural language prompt, we utilize a parameter-frozen text encoder to extract a semantic representation, followed by a linear layer for mapping its dimension to fit the transformer. To explore the impact of different text representations on style controlling, we attempt three types of encoders in our experiments: 1) BERT (Devlin et al., 2018), a widely-used self-supervised text encoder trained with masked-language modeling; 2) FLAN-T5 (Chung et al., 2022), the encoder of a unified text-to-text transformer fine-tuned with instructions; and 3) CLAP (Wu et al., 2023), a text encoder through contrastive pretraining on natural language and audio. We compare BERT and FLAN-T5 of different sizes, as well as CLAP pretrained on two different datasets. We also fine-tune BERT-large and FLAN-T5-large using prompts and corresponding labels. We fine-tune BERT with multi-label prediction and have FLAN-T5 predict the label sequence corresponding to the prompt in a text-to-text manner. Note that the prompts used in the evaluation are not included in fine-tuning.

4.3 Decoupled Pitch Representation

According to equal temperament theory(Wikipedia, 2023), humans' perception of musical intervals corresponds to the logarithmic distance of frequencies. This means if we multiply the fundamental frequency (F0) of the voiced part of singing by a factor (equivalent to adding an offset in the logarithmic domain), we can adjust the vocal range without changing the melody. Based on this principle, we decompose F0 into two components: 1) $\bar{f}_0$, which is the average value of the voiced part of F0, indicting the vocal range; and 2) $\tilde{f}_0$, where we rescale the voiced part of the original F0 sequence to have a specific mean value (230Hz, in our practice), indicating vocal-range-invariant melody information. This simple yet effective representation creates an information bottleneck, forcing the model to extract melodic and vocal range information from the rescaled F0 sequence and average F0 factor, respectively.

In our practice, we round $\tilde{f}_0$ and $\bar{f}_0$ into integers, and use an LUT to embed them before feeding them to the transformer backbone. Both $\tilde{f}_0$ and $\bar{f}_0$ share the same embedding space.

4.4 Alleviating Data Scarcity

Considering that both speech and singing are human voices in different forms, it is intuitive that they share some commonalities in style characteristics and distributions. Based on this point, we incorporate text-to-speech (TTS) data into the training of the prompt SVS task to alleviate data scarcity. Specifically, we employ the same methods as for singing to phonemize the text and generate prompts, and use an off-the-shelf tool to extract pitch from the speech, finally obtaining data items in the same format as SVS data.

Furthermore, we explore the feasibility of substituting speech data for singing data in low-resource scenarios. We evaluate the model performance under compositions of varying amounts of low-resourced SVS data with abundant TTS data, with experiment results presented in Section 5.5.

4.5 Multi-Scale Transformer Architecture

The multi-scale transformer serves as the backbone of our model. It is a decoder-only transformer with a hierarchical structure to facilitate the modeling of long sequences. This module aims to generate discrete acoustic units of singing voices conditioned on natural language prompts, lyrics phonemes, phoneme durations and vocal-range agnostic melody representation, together with the vocal-range factor as intermediate output. During training, the conditional inputs and target outputs are concatenated into a single sequence and fed to the transformer, which models the correlation using next-token-prediction with cross-entropy loss calculated on the target output part. During inference, the model predicts the range factor and acoustic units conditioned on the prefix input sequence autoregressively, which can be formulated as:

$$P_{cond}\left(\mathbf{a}\right) = P_{cond}\left(\bar{f}_0\right) \cdot \prod_{t=1}^{T} \prod_{c=1}^{C} P_{AR}\left(\mathbf{a}_t^c\right) \tag{1}$$

$$P_{cond}\left(*\right) = p\left(* \mid \mathbf{E}_P(P), L, D, \tilde{f}_0; \theta_{AR}\right) \tag{2}$$

$$P_{AR}\left(\mathbf{a}_t^c\right) = p\left(\mathbf{a}_t^c \mid \mathbf{a}_{<t}, \mathbf{a}_t^{<c}, \mathbf{E}_P(P), L, D, \tilde{f}_0, \bar{f}_0; \theta_{AR}\right) \tag{3}$$

where $\mathbf{a}$, $\mathbf{E}_P$, P, L, D, $\bar{f}0$, $\tilde{f}_0$ and θ_{AR} indicate acoustic units, prompt encoder, prompt, lyrics, durations, vocal-range factor, melody representation and model parameters, respectively, and t, c indicate temporal and codebook indices of the acoustic unit. Consider the process of the transformer predicting the vocal range factor, which is formulated by

$$P_{cond}\left(\bar{f}_0\right) = p\left(\bar{f}_0 \mid \mathbf{E}_P(P), L, D, \tilde{f}_0; \theta_{AR}\right), \tag{4}$$

as we assume that the average F0 value is independent of the lyrics, duration and melody, this formula indicates our model's capability to control the vocal range through natural language prompts. The predicted vocal range information is further taken as a condition for singing acoustic unit generation.

The hierarchical structure of the multi-scale transformer is illustrated in Figure 2(b). Such a hierarchical 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 acoustic units 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 $h_{1:t}$. 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 acoustic units of different codebooks inside a frame autoregressively. Such a design derives from a recent audio generation model (Yang et al., 2023b) and aims to reduce the computational complexity for over-long sequences caused by multi-codebook acoustic units. For non-acoustic modalities, each item is repeated n_q times to fit this modeling mechanism, with n_q being the number of codebooks.

4.6 Unit Vocoder

When the acoustic unit generation finishes, the generated units need to be mapped to a high-fidelity audio waveform. Due to the compressive nature of the codec, reconstructing audio from acoustic units of limited codebooks with the decoder may result in degraded perceptual quality. Instead of using the codec decoder directly, we adopt a GAN-based unit vocoder for singing voice reconstruction, aiming to generate audio of higher quality and richer details. Specifically, our vocoder is derived from BigVGAN (Lee et al., 2022), with a generator built from a set of look-up tables (LUT) that embed the discrete units, and a series of blocks composed of transposed convolution and a residual block with dilated layers. Multi-period and multi-resolution

discriminators (MPD, MRD) are used for adversarial training.

5 Experiments

5.1 Datasets

We combine 4 SVS datasets for our task, including M4Singer, Opencpop, Opensinger and PopCS, forming a multi-singer singing dataset of 127 hours. For speech data, we utilize 4 Mandarin TTS corpora, including AISHELL-3, Biaobei, THCHS-30 and a subset of DidiSpeech, totaling approximately 179 hours. We provide details of these datasets in Appendix B.

We phonemize the lyrics with PyPinyin[1], and extract F0 from raw audios with harvest (Morise et al., 2017). We separately select 2% of the singing data randomly for validation and testing, with the remaining used for training.

5.2 Model Configurations

The global transformer has 20 layers with 320M parameters, while the local transformer has 6 layers with 100M parameters. Both of them share the same hidden dimension of 1152. For acoustic units, we train a SoundStream model for 24k audio, with 12 quantization levels, a codebook size of 1024 and a downsampling rate of 480. We use the first 3 quantization levels as the acoustic units, and the unit vocoder is trained to reconstruct 24k audios from acoustic units of 3 codebooks. The label dropping probability p_1 and p_2 are both set to 0.25. Detailed structure and hyper-parameters of the model are appended in Appendix C.

5.3 Experiment Settings

As we are investigating a new task with no previous work to compare with, our experiments mainly focus on exploring different settings within our framework, including different text representations and different training data compositions, together with ablation studies. The settings of various text representations are presented in table 1. As described in Section 4.2, we experimented with encoders of different types, parameter sizes, and pre-training data as well as fine-tuning the encoders. We also provide the results of ground truth and two non-controllable SVS models in table 1 as baselines of singing quality: 1) FFT-Singer, which generates mel-spectrograms through stacked feed-forward transformer blocks; and 2) Diffsinger(Liu et al., 2022), an SVS model based on the diffusion probabilistic model.

In table 2, we compare the results of incorporating speech data for training or not, together with a series of low-resource data configurations with SVS data varying from 10 minutes to 100 hours paired with speech data of a fixed quantity of 100 hours. The ablation studies are described in a dedicated subsection.

5.4 Metrics

We employ both subjective and objective metrics to measure the controlling ability and singing voice quality of the models. For objectives metrics, we calculate the percentage accuracy for each attribute, where we train a gender classifier and use amplitude RMS and average F0 of the voiced part for volume and range evaluation. We mainly use single-attribute prompts for evaluation with an additional gender attribute for vocal range, and multi-attribute evaluation is conducted in ablation studies. We also calculate R-FFE for melodic accuracy between the synthesized and reference singing, which is F0-frame-error (FFE) with the voiced part of F0 rescaled to have an average of 230Hz to eliminate the impact of vocal range. For subjective metrics, we use crowd-sourced human evaluation via Amazon Mechanical Turk, where raters are asked to rate scores on 1-5 Likert scales on singing voice quality and the relevance between synthesized singing and the prompt. We report the mean-opinion-scores of quality (MOS) and relevance(RMOS) with 95% confidence intervals (CI) in the tables. Details of evaluation metrics are provided in Appendix D.

5.5 Results and Analysis

We can draw two basic conclusions from the results in table 1: 1) Generally, our models (1-10) exhibit favorable attribute controlling accuracies, with the best values being 87.7 / 86.3, 94.9 and 84.7 for the three attributes, together with competitive audio quality and melodic accuracy to non-controllable baselines (1-10 v.s. 11-13), with the best R-FFE and MOS being 0.09 and 3.90. This indicates the effectiveness of our model design on the task of controllable SVS. 2) The accuracies on volume are higher than gender and vocal range by a salient margin, with the values varying between 7.4 and 15.4 across different models. We speculate that this is because the random amplitude scaling in training allows the data with different volumes to be expanded to a large scale (somewhat similar to

[1]https://github.com/mozillazg/python-pinyin

ID	Model	Gender (F/M)	Volume	Range	R-FFE	MOS	RMOS
Prompt-Singer with Pre-trained Text Encoders							
1	FLAN-T5 small	76.7 / 78.1	92.0	79.1	0.11	3.75 ± 0.08	3.27 ± 0.09
2	FLAN-T5 base	82.2 / 79.5	92.4	80.8	0.12	3.79 ± 0.07	3.39 ± 0.07
3	FLAN-T5 large	83.1 / 80.8	92.7	82.6	0.12	3.83 ± 0.08	3.43 ± 0.08
4	FLAN-T5 XL	83.4 / 80.4	92.6	82.9	0.11	3.84 ± 0.06	3.46 ± 0.08
5	BERT-base	80.8 / 80.1	93.9	80.1	0.10	3.81 ± 0.06	3.42 ± 0.07
6	BERT-large	84.9 / 80.9	94.3	78.9	**0.09**	3.78 ± 0.08	3.44 ± 0.08
7	CLAP-general	82.2 / 79.5	94.1	80.3	0.12	3.83 ± 0.07	3.43 ± 0.06
8	CLAP-speech/music	82.2 / 78.1	94.2	80.8	0.11	3.85 ± 0.09	3.38 ± 0.08
Prompt-Singer with Fine-tuned Text Encoders							
9	FLAN-T5 large finetuned	**87.7 / 86.3**	94.4	**84.7**	0.12	3.89 ± 0.07	$\mathbf{3.62 \pm 0.08}$
10	BERT-large finetuned	86.3 / 83.6	**94.9**	79.8	0.10	$\mathbf{3.90 \pm 0.07}$	3.60 ± 0.08
Non-controllable SVS models and Ground Truth							
11	FFT-Singer	/	/	/	0.17	3.67 ± 0.08	/
12	Diffsinger	/	/	/	**0.09**	3.86 ± 0.07	/
13	Ground Truth	98.0 / 97.0	/	/	/	4.09 ± 0.06	/

Table 1: Results on different text representations, including percentage accuracies of the three attributes, rescaled f0-frame error (R-FFE) and mean-opinion-scores of audio quality (MOS) and relevance to the prompt (RMOS).

ID	SVS Data	TTS Data	Gender (F/M)	Volume	Range	R-FFE	MOS	RMOS
1	✓	✗	75.3 / 65.8	87.6	78.7	0.11	3.68 ± 0.08	3.37 ± 0.08
2	✓	✓	87.7 / 86.3	94.4	84.7	0.12	3.89 ± 0.07	3.62 ± 0.08
3	10min	100h	65.8 / 65.6	78.3	80.9	0.29	3.06 ± 0.09	2.89 ± 0.09
4	1h	100h	71.2 / 64.4	84.8	81.2	0.25	3.34 ± 0.08	3.03 ± 0.09
5	10h	100h	76.7 / 68.5	88.6	81.6	0.23	3.28 ± 0.08	3.17 ± 0.09
6	100h	100h	86.2 / 80.5	92.5	82.3	0.12	3.75 ± 0.08	3.45 ± 0.08

Table 2: Experiment results on data scarcity alleviation in low resource scenarios.

data augmentation), while the quantities and diversities of gender and range are limited by the training datasets. This, from one perspective, confirms that data scarcity makes learning the correlation between prompt and style attributes difficult.

5.5.1 Evaluation on Text Representations

We have the following further observations from the results in table 1: 1) Fine-tuning the text encoders leads to a considerable improvement in controlling accuracy (3 vs. 9 & 6 vs.10), with the improvements being 4.6 / 5.5, 1.7 and 2.1 for FLAN-T5 large, and 1.4 / 2.7, 0.6 and 0.9 for BERT-large. This indicates that aligning the text representations with the labels, which have a much simpler distribution, helps the model learn their correlation with singing style. Nevertheless, using only the pre-trained text encoders already yields quite good results. 2) Generally, larger model sizes bring better results (1-4 & 5-6). However, such a tendency between 3 and 4 is less significant compared to 1-2 and 2-3, suggesting that text encoder parameters beyond a certain size are no longer a bottleneck for model performance. 3) Different types of text encoders exhibit varying controlling capabilities over different attributes. For instance (1-4 vs. 5-8), the FLAN-T5 family shows weaker control over volume compared to CLAP and BERT, with an accuracy gap of 1.2-2.3. However, the large and xl models outperform CLAP and BERT in vocal-range controlling accuracy by 1.8-4.0. This may be related to differences in the models' pretraining methods and data. We choose the fine-tuned FLAN-T5 large model for subsequent experiments.

5.5.2 Evaluation on Data Scarcity Alleviation

From the results of different data compositions in table 2, we have the following observations: 1) Introducing speech data leads to a comprehensive improvement in controlling accuracies and generation quality, with the cost being a slight increase in R-FFE of 0.01 (1 vs. 2). This is because the additional speech data increases the quantity and diversity of the training data, aiding the network in modeling the correlation between prompt and acoustic style. However, due to the difference in the

ID	Model	Gender (F/M)	Volume	Range	R-FFE	RMOS
Ablation on Decoupled Pitch Representation						
1	Factor: ✓ Rescale: ✓	87.7 / 86.3	94.4	84.7	0.12	3.62 ± 0.08
2	Factor: ✗ Rescale: ✓	78.1 / 63.0	91.3	76.1	0.11	3.34 ± 0.09
3	Factor: ✗ Rescale: ✗	64.4 / 58.9	91.6	72.3	0.08	2.75 ± 0.09
Ablation on Different Prompted Attribute Numbers						
4	Attribute Num: 1	87.7 / 86.3	94.4	/	0.12	3.67 ± 0.08
5	Attribute Num: 2	84.3 / 82.9	93.4	84.7	0.11	3.58 ± 0.08
6	Attribute Num: 3	81.2 / 80.7	93.0	82.4	0.11	3.52 ± 0.07

Table 3: Results of ablation studies.

distributions of singing melody and speech prosody, both of which are manifested in pitch variation, the speech data may have a negative impact on modeling singing melody, causing the slight increase in R-FFE. 2) In the low resource scenarios (3-6), we find that there is a drastic decline in the singing audio quality, melody accuracy as well as the accuracy on gender with the decrease in the quantity of SVS data. In contrast, the changes in volume and vocal range are relatively gradual, yielding acceptable results of 88.6 and 81.6 even with 10 hours of singing data. This suggests that, while speech data helps improve controlling accuracy and audio quality, it still cannot substitute for singing data in modeling certain vocal characteristics. In conclusion, introducing speech data effectively enhances the performance of controllable SVS, but it is still necessary to have a sufficient amount of singing data to ensure synthesis quality and melody accuracy.

5.6 Ablation Studies

We mainly focus on validating the effectiveness of our decoupled pitch representation and multi-attribute prompting mechanism in the ablation studies, and the results are presented in table 3.

For pitch representation (1-3), we first remove the vocal range factor from the sequence, and then eliminate the rescaling on the input F0. We can see that when removing the range factor, there is a drastic drop of 9.6 / 23.3, 3.1 and 8.6 in accuracies, accompanied by an RMOS decrease of 0.28. This indicates that explicitly predicting the vocal range factor facilitates vocal range and gender control greatly. When we continue to eliminate the input F0 rescaling, the accuracies on gender and range as well as RMOS further decline by 13.7 / 4.1, 3.8 and 0.59, respectively, which indicates that the vocal range information contained in the original F0 interferes with the model's modeling of the correlation between prompt and singing style. We also observe that removing the range factor and input F0 rescaling leads to an improvement in melodic accuracy. This suggests that the decoupling mechanism may cause some loss of pitch information. Despite this, our model keeps a satisfactory melodic accuracy with the decoupled pitch representation.

We further examine the model's controlling effectiveness under multi-attribute prompts. The results of 4-6 in table 3 show that there is a slight decrease in accuracies and RMOS as the attribute number increases, with the drop being 3.4 / 3.4, 1.0, 0.09 from 1 to 2 attributes, and 3.1 / 2.2, 0.4, 2.3, 0.06 from 2 to 3. We suggest that this is because the conditional distribution of acoustic style with respect to controlling signals of multiple attributes is more complicated to be modeled. Nevertheless, our model shows favorable performance on prompts with both single and multiple attributes.

6 Conclusion

In this paper, we propose Prompt-Singer, the first singing-voice-synthesis method with the ability of style control using natural language prompts. We adopt a multi-scale decoder-only transformer for generating acoustic units of singing, followed by a unit-vocoder for audio reconstruction. We design a decoupled pitch representation for vocal range modification with an accurate melody kept. Furthermore, we investigate various experiment settings, including different text representations, fine-tuning the text encoders, and using speech data to boost performance in low-resource scenarios.

In future works, we plan to introduce more style attributes in controllable SVS, such as emotion, rhythm and more detailed singer information. We hope our work will facilitate the development of the SVS community.

7 Limitations and Potential Risks

Despite that our model achieves remarkable controlling capability and audio quality on prompt singing-voice-synthesis, it still has two major limitations: 1) Our descriptions of singing voice styles are based on a limited set of predefined categories, which constrains our model's ability to control specific acoustic attributes to a coarse granularity. 2) Due to the limitation of text generation capability of the LLM used for prompt generation, the generated prompts may suffer from lower diversity compared with real-world instructions, and may have a bias in distribution, which may limit the potential of real-world applications of our model. Besides, misuse of our model for singing voice generation may lead to 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

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jiawei Chen, Xu Tan, Jian Luan, Tao Qin, and Tie-Yan Liu. 2020. Hifisinger: Towards high-fidelity neural singing voice synthesis. *arXiv preprint arXiv:2009.01776*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhiyong Zhang Dong Wang, Xuewei Zhang. 2015. Thchs-30 : A free chinese speech corpus.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023a. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. 2023b. Natural language supervision for general-purpose audio representations.

Tingwei Guo, Cheng Wen, Dongwei Jiang, Ne Luo, Ruixiong Zhang, Shuaijiang Zhao, Wubo Li, Cheng Gong, Wei Zou, Kun Han, et al. 2021. Didispeech: A large scale mandarin speech corpus. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6968–6972. IEEE.

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.

Rongjie Huang, Feiyang Chen, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. 2021. Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3945–3954.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2022. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*.

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*.

Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. 2022. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11020–11028.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Masanori Morise et al. 2017. Harvest: A high-performance fundamental frequency estimator from speech signals. In *INTERSPEECH*, pages 2321–2325.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and

Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2020. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567*.

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

Yu Wang, Xinsheng Wang, Pengcheng Zhu, Jie Wu, Hanzhao Li, Heyang Xue, Yongmao Zhang, Lei Xie, and Mengxiao Bi. 2022. Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis. *arXiv preprint arXiv:2201.07429*.

Wikipedia. 2023. Equal temperament — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Equal%20temperament&oldid=1185261297. [Online; accessed 01-December-2023].

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Guangzhi Lei, Chao Weng, Helen Meng, and Dong Yu. 2023a. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *arXiv preprint arXiv:2301.13662*.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, et al. 2023b. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*.

Jixun Yao, Yuguang Yang, Yi Lei, Ziqian Ning, Yanni Hu, Yu Pan, Jingjing Yin, Hongbin Zhou, Heng Lu, and Lei Xie. 2023. Promptvc: Flexible stylistic voice conversion in latent space driven by natural language prompts. *arXiv preprint arXiv:2309.09262*.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

Lichao Zhang, Ruiqi Li, Shoutong Wang, Liqun Deng, Jinglin Liu, Yi Ren, Jinzheng He, Rongjie Huang, Jieming Zhu, Xiao Chen, et al. 2022a. M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. *Advances in Neural Information Processing Systems*, 35:6914–6926.

Yongmao Zhang, Jian Cong, Heyang Xue, Lei Xie, Pengcheng Zhu, and Mengxiao Bi. 2022b. Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7237–7241. IEEE.

Yongmao Zhang, Guanghou Liu, Yi Lei, Yunlin Chen, Hao Yin, Lei Xie, and Zhifei Li. 2023a. Promptspeaker: Speaker generation based on text descriptions. *arXiv preprint arXiv:2310.05001*.

Zewang Zhang, Yibin Zheng, Xinhui Li, and Li Lu. 2022c. Wesinger: Data-augmented singing voice synthesis with auxiliary losses. *arXiv preprint arXiv:2203.10750*.

Zewang Zhang, Yibin Zheng, Xinhui Li, and Li Lu. 2023b. Wesinger 2: fully parallel singing voice synthesis via multi-singer conditional adversarial training. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

A Sample Prompt Keywords and Sentence Templates

We list the keywords for each category in table 4, and provide some samples of prompt sentence templates in table 6.

Category	Keywords
Gender	
female	woman, lady, girl, female, lass, miss, madam
male	man, boy, guy, gentleman, male, sir
Volume	
high	loud, ringing, booming, thunderous, deafening, roaring
medium	moderate, average, intermediate, middle-range
low	quiet, slight, twittering, hushed, whispering
Vocal Range	
high	sharp, treble, shrill, whistling, shrieking, high-pitched
low	deep, low, bass, thick, low-pitched

Table 4: Prompt keywords for each category.

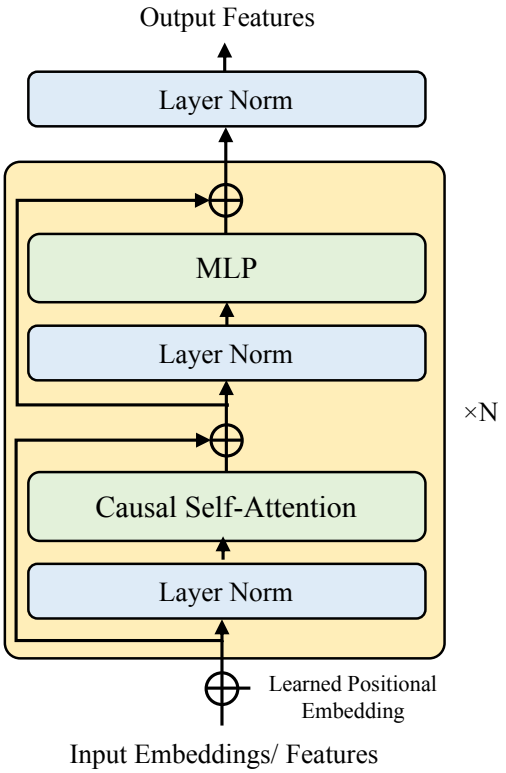


Figure 3: Structure of Global Transformer

B Dataset Statistics

In table 5, we list the statistics of the datasets used. F and M in the Speakers column indicate the numbers of female and male speakers or singers.

Dataset	Hours	Speakers
SVS datasets		
M4Singer (Zhang et al., 2022a)	29.8	F:10 M:10
Opencpop (Wang et al., 2022)	5.3	F:1
Opensinger (Huang et al., 2021)	86.5	F:49 M:28
PopCS (Liu et al., 2022)	5.9	F:1
TTS datasets		
AISHELL-3 (Shi et al., 2020)	86.4	F:176 M:42
Biaobei [2]	11.8	F:1
THCHS-30 (Dong Wang, 2015)	34.2	F:31 M:9
Didispeech (Guo et al., 2021)	47.0	F:198 M:202

Table 5: Statistics of training datasets.

C Model Settings

We illustrate the architecture of the global transformer in Figure 3. The local transformer shares the same structure as the global one with two differences: 1) the local transformer has no positional embedding, and 2) there is a linear lm-head appended to the top of it for token prediction. We also list the model hyper-parameters of Prompt-Singer in Table 7. The multi-scale transformer is trained with 6 NVIDIA-V100 gpus for about 4-5 days, and the vocoder is trained with 4 NVIDIA-V100 gpus for a week.

D Evaluation Metrics

D.1 Objective Evaluation

For gender controlling accuracy, we train an open-source gender classifier[3] with our singing and speech data. The performance of the classifier on the test set is provided as ground-truth accuracy in line 13 of table 1.

For controlling accuracies on volume and vocal range, considering that the values of generated singing may slightly deviate from the boundaries used for categorization, we adopt a soft-margin mechanism for accuracy calculation. Specifically, we take the accuracy of data falling within the correct range as 100, and calculate the accuracy with $100 * \exp(-k\epsilon)$ for data outside the correct range, where ϵ is the error between the data value and the boundary, and k is a hyper-parameter controlling the decay rate of accuracy at the margins, with

[2] https://www.data-baker.com/open_source.html

[3] https://github.com/x4nth055/gender-recognition-by-voice/tree/master

Single-Attribute Templates
Do you have any songs with a [gender] lead singer?
Can you create a song sung by a [gender] vocalist?
I'm searching for a song featuring a [gender] singer.
I need a song with a [volume] voice that resonates.
Play me a song with a [volume] voice.
I'd like to listen to a song with a [volume] voice.
I need a song where every note is gentle and delicate. (for low volume)
Kindly provide me with a song that features a voice of balanced volume, pleasing to the ears. (for medium volume)
Give me a song with a voice that shakes the ground with its thunderous vocals! (for high volume)
Double-Attribute Templates
Can you find me a song with a [gender] singer and a [volume] voice?
I would like to hear a song with a [volume] voice and if possible, a [gender] voice.
Synthesize a new song with a [volume] voice and a [gender] lead singer.
Need a [pitch] pitch song sung by a [gender] vocalist.
Generate a song featuring a [gender] vocalist with a unique use of [pitch] pitch.
A [gender] voice with a [pitch] pitch is what I'm looking for.
Create an enchanting song sung by a [gender] vocalist in the [pitch] pitch.
Create a [gender] artist's song with a [volume] voice, softly mesmerizing with its gentle tone. (for low volume + any gender)
Generate a [gender] artist singing at just the right volume. (for medium volume + any gender)
Can you generate a [gender]-sung song with a [volume] voice that balances softness and loudness? (for medium volume + any gender)
I'm looking for a song with a [gender] singer and a voice that's as powerful as a thunderstorm. (for high volume + any gender)
Triple-Attribute Templates
Explore [gender] [volume] songs with emotive [pitch] pitch.
Synthesize a song with a [pitch] pitch and a [volume] voice, preferably [gender].
Design a [gender] singer's song with a [volume] voice and [pitch] pitch.
Showcasing superb [pitch] pitch, create a [volume] song by a [gender] artist.
Generate a song with stunning [pitch] harmonies and a [gender] singer with a [volume] voice.
Can you compose a song with a [gender] vocalist and [volume] volume, while incorporating the singer's unique use of [pitch] pitch?
Generate a song featuring [gender] vocals, delicately whispered with [volume] voice and [pitch] harmony. (for low volume + any gender / vocal range)
Compose a [pitch]-keyed song with a [volume] voice that balances softness and loudness, sung by a [gender] singer. (for medium volume + any gender / vocal range)
Craving a [gender] artist's song with a [volume] voice that exudes energy and power and a [pitch] note that creates a memorable hook! (for high volume + any gender / vocal range)

Table 6: Sample Prompt Sentence Templates.

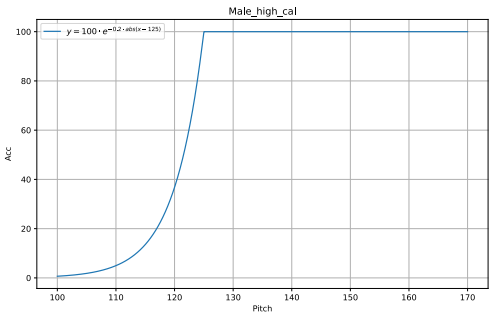


Figure 4: Soft-margin accuracy curve of high vocal-range of male.

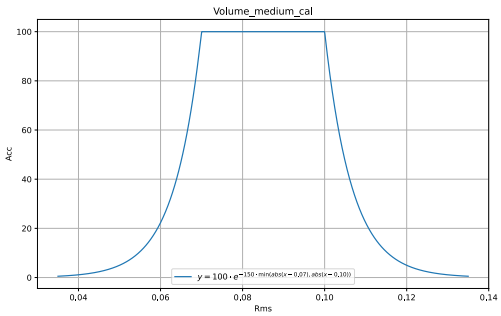


Figure 5: Soft-margin accuracy curve of medium volume.

larger k corresponding to faster decay. We take accuracy curves of high vocal-range of male and medium volume as examples and illustrate them in Figure 4 and 5, respectively. We set k to 120, 150 and 180 for high, medium and low volume, and 0.2 for vocal range accuracy.

D.2 Subjective Evaluation

For each evaluated model, we mix all generated results together and randomly select 220 items with

Hyperparameter		Prompt-Singer
Global Transformer	Layers	20
	Hidden Dim	1,152
	Attention Headers	16
	FFN Dim	4,608
	Number of Parameters	320.07M
Local Transformer	Layers	6
	Hidden Dim	1,152
	Attention Headers	8
	FFN Dim	4,608
	Number of Parameters	100.13M
Unit Vocoder	Upsample Rates	[6,5,2,2,2,2]
	Hop Size	480
	Upsample Kernel Sizes	[12,9,4,4,4,4]
	Number of Parameters	125.43M

Table 7: Hyperparameters of Prompt-Singer.

their corresponding prompts for subjective evaluation.

Our subjective evaluation tests are crowdsourced and conducted via Amazon Mechanical Turk. For audio quality evaluation, we ask the testers to examine the audio quality and naturalness and ignore the content. For prompt-style relevance, we instruct the testers to evaluate the relevance between the natural language prompt and the singing style while ignoring the content. The testers rate scores on 1-5 Likert scales. We provide screenshots of the testing interfaces in Figure 6 and 7. Each data item is rated by 4 testers, and the testers are paid $8 hourly.

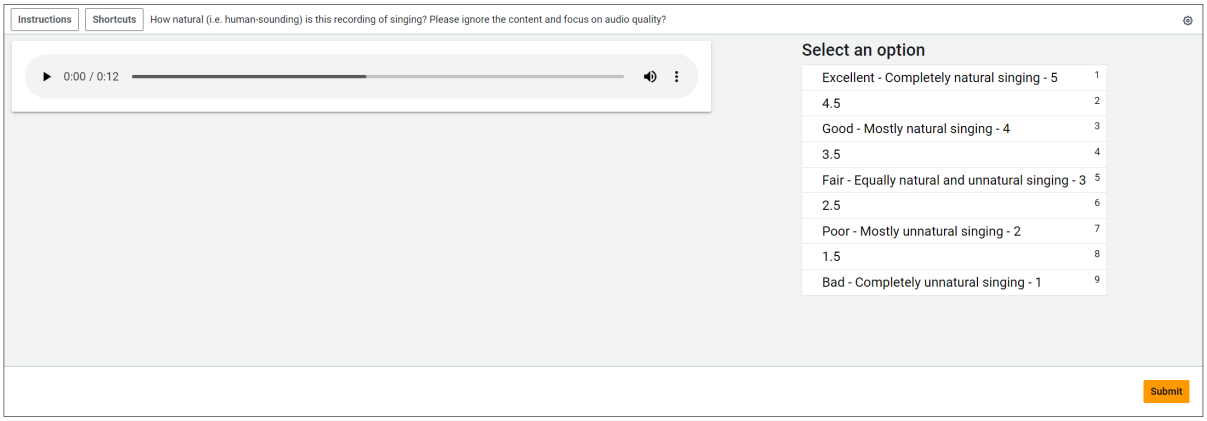


Figure 6: Screenshot of MOS testing.

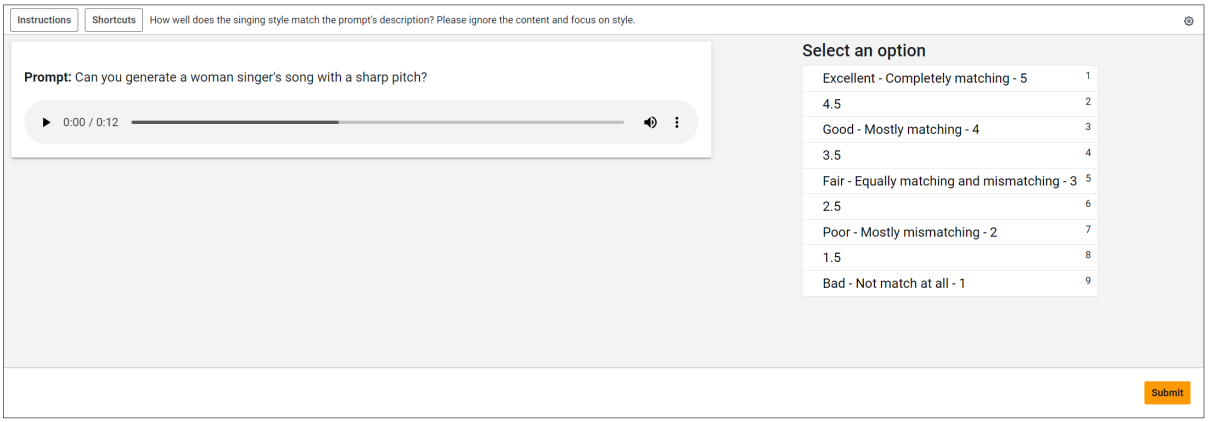


Figure 7: Screenshot of RMOS testing.