Text-to-Song: Towards Controllable Music Generation Incorporating Vocals and Accompaniment

Anonymous ACL submission

Abstract

A song is a combination of singing voice and accompaniment. However, existing works focus on singing voice synthesis and music generation independently. Little attention was paid to explore song synthesis. In this work, we propose a novel task called text-to-song synthesis which incorporating both vocals and accompaniments generation. We develop Melodist, a two-stage text-to-song method that consists of singing voice synthesis (SVS) and vocal-to-accompaniment (V2A) synthesis. Melodist leverages tri-tower contrastive pretraining to learn more effective text representation for controllable V2A synthesis. A Chinese song dataset mined from a music website is built up to alleviate data scarcity for our research. The evaluation results on our dataset demonstrate that Melodist can synthesize songs with comparable quality and style consistency. Audio samples can be found in `https://text2songMelodist.github.io/Sample/`.

1 Introduction

Songs, as intricate musical compositions, have always enjoyed the greatest popularity among music enthusiasts. It inspires the pursuit of song synthesis by leveraging machine learning and artificial intelligence algorithms. It makes sense to generate a song conditioned on text modality (music score, natural language prompt, etc.). However, there is little exploratory research on text-to-song synthesis to our knowledge.

There are two related tasks. The first is singing voice synthesis, which converts the music score (lyrics, notes, and duration) to the singing voice. Existing SVS models have achieved remarkable achievement regarding quality(Huang et al., 2021; Liu et al., 2022; Hong et al., 2023; Zhang et al., 2022a) and zero-shot ability(Qian et al., 2019; Casanova et al., 2022) but they can only generate vocals. Another similar task is accompaniment generation(Ren et al., 2020; Dong et al., 2018; Ding and Cui, 2023), which usually aims at generating multi-track sequences for melody or a given lyrics(Madhumani et al., 2020; Yu et al., 2021) in the symbolic domain. As presented in Figure 1, there are similarities among these three tasks, while notable distinctions exist. The accompaniments are often removed in data preprocessing to train a SVS model. And existing music generation models do not take vocals into account as the condition. Further exploration of text-to-song is inhibited.

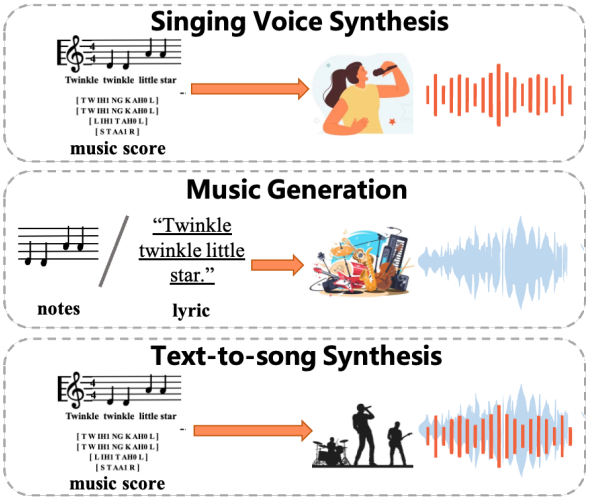

Figure 1: The comparison of three tasks: singing voice synthesis, accompaniment generation and text-to-song. In this work, We investigate on the relationship between vocal and accompaniment for text-to-song synthesis.

Neither serves as the suitable prior. To address this limitation, we propose a novel generative task, *Text to Song*, which converts the music score (lyrics, notes, and duration) to the song, that is, singing voice with accompaniment. However, a text-to-song model is facing several challenges:

1) Process of Synthesis. It is hard to achieve end-to-end generation since the song contains much more information (pitch variation, timbre,

emotion, instruments, etc.) than the music score, which imposes a large burden on the model.

2) Attribute Controllability. This is far from enough to model the diverse output while only feeding the music score to the text-to0song synthesis model. Some natural language prompts should be included as the condition to guide and control the accompaniment generation.

3) Data Scarcity. To the best of our knowledge, there is no dataset with pairs of vocal and accompaniment audios along with finely annotated music score (which should at least have lyrics transcription). It is the most intractable factor hindering research in this area.

In this paper, we propose Melodist, the first text-to-song model to generate music incorporating vocals and accompaniments from music score. To overcome the challenges mentioned above, we adopt several techniques: 1) Based on the insight(Défossez et al., 2019) that the song signal can be naively regarded as the mixture of accompaniment signal and vocal signal, we introduce a two-stage text-to-song synthesis. Specifically, Melodist generates singing voice from the music score in Stage 1, then generates accompaniment given vocals in Stage 2. Finally, we mix the outputs of two stages to obtain the song. It releases the burden of our model to a large extent; 2) We utilize the attribute tags (mood, instruments, style, etc.) of each song segment and construct natural language prompts to control the synthesis of the accompaniment. We further apply the Tri-Tower Contrastive Learning framework to extract better text representations; 3) We crawled some songs and the corresponding lyrics and tags related to attributes from music websites. We evaluate our model under different settings and the results demonstrate that Melodist can synthesize songs with comparable quality and attribute controllability.

The main contributions of our work can be summarized as follows:

- We introduce a new task of text-to-song synthesis, which aims to convert the music score to the song incorporating vocal and accompaniment synthesis. We further propose Melodist, the first text-to-song model following two-stage song synthesis;
- We adopt natural language prompts to generate various types of accompaniment for the purpose of attribute controllability;
- We design a tri-tower contrastive learning framework to connect the text context with its corresponding vocal and accompaniment pattern;
- We construct a dataset that provides not only pairs of vocals and accompaniment but also transcriptions in text format including lyrics and attribute tags.
- We conduct extensive experiments to verify the effectiveness of Melodist. Experiment results show that Melodist exhibits high quality and great adherence.

2 Related Work

2.1 Singing Voice Synthesis

Substantial progress has been made in Singing Voice Synthesis (SVS). Several works(Huang et al., 2022b; Kong et al., 2020) have adopted generative adversarial networks (GANs)(Goodfellow et al., 2020), while there appear many end-to-end SVS models(Zhang et al., 2022b; Hong et al., 2023) based on variational autoencoder (VAE). DiffSinger(Liu et al., 2022) is built on diffusion probabilistic models which can generate more high-fidelity outputs. In the realm of the Large Language Model recently, there are many emerging methods(Yang et al., 2023; Huang et al., 2023) modeling voice with an auto-regressive transformer in a compact and discrete space. However, these works discarded the accompaniments in data preprocessing, while we take accompaniment generation into account and investigate the relationship between vocals and accompaniments.

2.2 Accompaniment Generation

Researchers on accompaniment usually work on musical symbolic tokens in a seq2seq setting. MuseGAN(Dong et al., 2018) is the first model that generate multi-track polyphonic music with harmonic and rhythmic. There exist several works (Copet et al., 2023; Agostinelli et al., 2023)trying to generate melody conditions on chord information for better music structure. Yang et al.(Yang et al., 2017) designed MidiNet to generate melodies one bar after another. PopMAG (Ren et al., 2020)was proposed to simultaneously generate five instrumental tracks in a single sequence. However, these methods rely highly on symbolic music representation. Recently, Donahue et al. presented SingSong(Donahue et al., 2023), a system that generates instrumental music to accompany input vocals. But the limitation remains in the lack of controllability related to mood, instruments, style, etc.

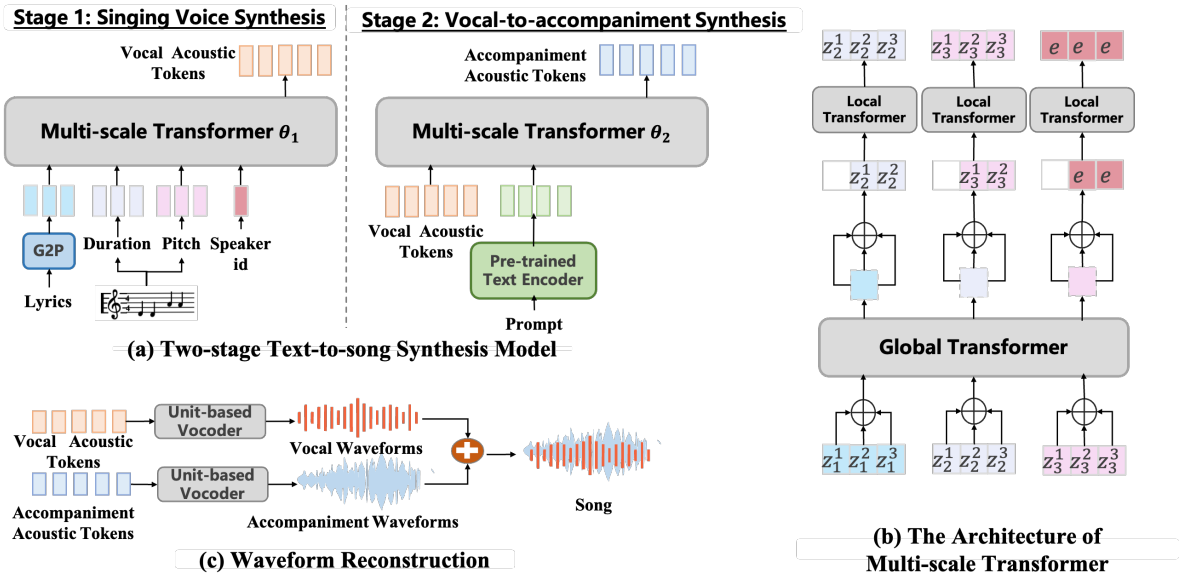

Figure 2: The overview of Melodist, the proposed two-stage text-to-song synthesis model. We present the two-stage pipeline in subfigure (a). In subfigure (b), we present the multi-scale Transformer architecture, in which e and z_t^k denote <EOS> token and the k-th audio token at t-th frame, respectively.

In this work, we focus on developing a controllable text-to-song synthesis model that accepts more user controls for guiding the generation results.

2.3 Cross-modal Contrastive Learning

Contrastive learning, which is first applied in computer vision domain(Radford et al., 2021; Oord et al., 2018), achieves high performance in many downstream tasks such as zero-shot recognition, image-text retrieval, etc. Along the same line in the audio domain, Wav2clip(Wu et al., 2022) and Audioclip(Guzhov et al., 2022) are both derived from CLIP. To achieve more flexibility and generalization, CLAP(Elizalde et al., 2023) is proposed to learn audio concepts from natural language supervision instead of class labels. Recently, an increasing number of works(Chen et al., 2022; Manco et al., 2022) exploring contrastive pre-training in the music domain. MuLan(Huang et al., 2022a) is the first model learning a joint embedding space for music and natural language trained with an unprecedented scale of weakly paired text and audio. In this work, we also leverage a contrastive learning framework to extract better text representations.

3 Two-stage Text-to-song Synthesis

In this section, we first present a formal Definition of text-to-song synthesis task. Then we will give an overview of the proposed model Melodist. Finally, we will elaborate on the approaches we adopt for controllable two-stage text-to-song synthesis.

3.1 Task Definition

In this work, we present a novel task *text-to-song* and extend it to controllable synthesis. Given the training set D consists of n data points (s_i, p_i, c_i), $i = 1, ..., n$, where each element denotes a song, the description of its accompaniment and music score of its vocal, we convert the music score to song conditioned on the natural language prompt, which can be formulated as a conditional probability distribution modeling problem:

$$p(S|C, P) = \prod_{t=0}^{T} p(s_t|s_{<t}, C, P; \theta) \qquad (1)$$

Given that $S = S_v + S_a$, where S_a, S_v, S_a denote song waveforms, vocal waveforms and accompaniment waveforms respectively, we can redefine *text-to-song* task as the approximation of joint conditional probability optimization $p(S_v, S_a|C, P)$, where S_v and S_a denote vocals and accompaniment, respectively.

3.2 Overview

In this work, we propose Melodist, the first controllable text-to-song model. As illustrated in Figure 2, it is organized in two stages: 1) In the first stage we follow the common SVS process that generates a singing voice conditioned on the music

score, and in the second stage we generate musical accompaniments from singing given natural language prompt. Instead of directly modeling distributions over vocal and accompaniment waveforms, we adopt acoustic tokens as the prediction targets. Finally, we reconstruct waveforms from predicted vocal acoustic tokens and accompaniment acoustic tokens and then mix them as the output.

Two-stage generation enjoys the following advantages: 1) It reflects the conditional independence assumption that the attribute control applied on accompaniment is independent of the vocals and music score; 2) It is consistent with the dependency that the semantic and acoustic features of singing voice depend on music score while the harmony and controllability of accompaniment are decided on vocals and prompts, respectively.

3.3 Predicted Target

Acoustic tokens, as the predicted target, are extracted the acoustic tokens by SoundStream(Zeghidour et al., 2021), a neural codec with an encoder-decoder architecture and a residual vector quantizer (RVQ) cascaded n_q layers of vector quantizer (VQ). Assuming y denotes an audio sample, the extracted acoustic tokens sequence can be represented as $Z^{Tn_q} = encoder(y)$ where T refers to the number of frames. These compressed representations can be used to reconstruct waveforms by the decoder subsequently that $\hat{y} = decoder(Z)$.

3.4 Backbone Model

We adopt the multi-scale transformer proposed in(Yu et al., 2023; Yang et al., 2023) as our backbone in both two stages. It introduces a hierarchical design consisting of a global transformer and a local transformer, both of which are decoder-only transformers. Specifically, the flattened acoustic token sequence is first chunk into patches $\{x_0, x_1, \ldots, x_T\}$ of T frames, each containing n_q tokens of one frame. The chunked sequence is passed to the global transformer G to predict the target in a frame-by-frame manner:

$$H^{g_out}_{1:T} = G(H^{g_in}_{0:T-1}), \tag{2}$$

In contrast, the local model L operates on a single patch of size n_q, each of which is the sum of the output of the global model and the embedding of the previous tokens.

$$H^{l_out}_{t,1:n_q} = L(WH^{g_out}_{t-1,0:n_q-1} + H^{l_in}_{t,0:n_q-1}) \tag{3}$$

Where W denotes the projection matrix to map the hidden size of the local transformer.

During training, the model is optimized using token prediction and cross-entropy loss. In the inference stage, the model autonomously predicts acoustic tokens in an auto-regressive manner conditioning on prefixed input sequences. Such a design facilitates the reduction of computational and enhances in-context learning for long sequences to a large extent.

3.5 Two-stage Synthesis

Stage 1: Singing Voice Synthesis. In the SVS stage, the model synthesizes acoustic tokens conditioned on lyric phonemes, phoneme durations, and phonemes pitch. Specifically, we transform the condition input into discrete tokens and repeat each for n_q times to fill each patch. The expanded inputs and target acoustic tokens are concatenated and embedded into a unified sequence, subsequently processed by the multi-scale transformer.

Stage 2: Vocal-to-accompaniment Synthesis. In the vocal-to-accompaniment synthesis stage, the model synthesizes acoustic tokens of accompaniment conditioned on vocal acoustic tokens and natural language prompts. We leverage a pre-trained text encoder providing text representation with consistent global characteristics with the vocal and accompaniment, which we will illustrate in section 4 in detail. It can be incorporated with our backbone model to enhance attribute controllability. We freeze the parameters of the text encoder, utilize it to extract the non-pooled text representation of the prompt, and pass it through a linear layer to fit the dimension of the backbone model. Once we have obtained "continuous text embeddings", we also repeated each token for n_q times. The inputs are organized and processed in the same way as in the previous stage.

3.6 Waveform Reconstruction

Instead of the decoder of Soundstream, we adopt a unit-based vocoder utilizing GAN-based architecture for waveform generation from acoustic units. It is derived from BigvGAN and comprises a generator and two discriminators. Specifically, the generator is built from a set of look-up tables (LUT) that embed the discrete units. It is followed by a series of blocks composed of transposed convolution for the purpose of upsampling and a residual block with dilated layers to expand the receptive

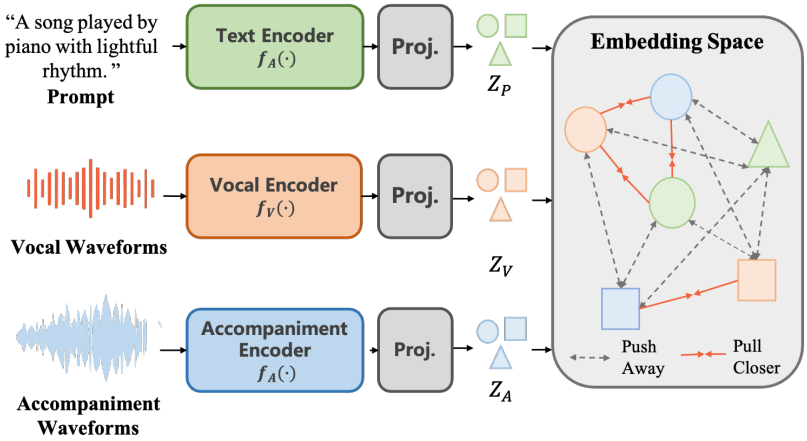

Figure 3: The architecture of the tri-tower contrastive framework. Z_P, Z_V, Z_A refer to the representation extracted by the text encoder, the vocal encoder and the accompaniment encoder, respectively. We use different shapes to represent different triples, while color is used to distinguish the kinds of inputs. Embeddings of the same triplet are pulled closer, while those of different objects are pushed away in the joint embedding space.

field. The multi-period discriminator (MPD) and the multi-resolution discriminator (MRD) proposed in BigvGAN are added to distinguish between the generated audio and ground truth. Note that we train two neural codecs (the vocoders and encoders used to extract acoustic tokens) sharing the same architecture but not the same parameters respectively for vocals and accompaniments. We found that gradient collapse occurs when training only one neural codec on all audios. It is mainly attributed to the distribution discrepancy between the vocals and accompaniments. Once we obtain the waveforms of the vocal and its accompaniments, we mix them in the waveform domain to get the final output.

4 Tri-Tower Contrastive Pre-training

We introduce a tri-tower training scheme with contrastive loss that jointly embeds text, vocals, and accompaniments into a feature-aligned space. As presented in Figure 3, it consists of three separate encoders: text encoder $f_P(\cdot)$, vocal encoder $f_V(\cdot)$, and accompaniment encoder $f_A(\cdot)$, each followed by a pooling and linear layer. Parallel text prompt, vocal, and accompaniment make up each triplet of a mini-batch (x_p, x_v, x_a) and they are passed through the respective encoder. The text encoder $f_P(\cdot)$: $\mathcal{A}^n \rightarrow \mathbb{R}^{d_P}$ converts a tokenized text sequence of length n over vocabulary A to the text embedding of dimension d. The Vocal encoder and the accompaniment encoder $f_V(\cdot)$, $f_A(\cdot)$: $\mathbb{R}^{F\times T} \rightarrow \mathbb{R}^{d_V}$ encode log mel spectrograms of the vocal and accompaniment respectively, which F refers to the number of mel channels and T refers to the number of frames. A linear layer is appended in each branch to project the representations into a l_2-normalized embedding space.

When considering two-tower contrastive learning, two encoders of different modalities are jointly trained to maximize the similarity between N positive pairs while minimizing the similarity for N $\times(N-1)$ negative pairs. We adopt the multi-modal version of InfoNCE loss(Oord et al., 2018). Taking pair (text, vocal) as an example, the loss can be formulated as follows:

$$L_{p\rightarrow v} = -\log \frac{\exp(z_{p_i} \cdot z_{v_i}/\tau)}{\sum_{j=1}^{N} \exp(z_{p_i} \cdot z_{v_j}/\tau)} \quad (4)$$

$$L_{p\leftrightarrow v} = (L_{p\rightarrow v} + L_{v\rightarrow p})/2 \quad (5)$$

Where τ is a temperature parameter. To extend it into tri-tower contrastive loss, we simply calculate the contrastive loss over pairs of the representations in a triplet (text, vocal, accompaniment) that:

$$L = L_{p\leftrightarrow v} + L_{p\leftrightarrow a} + L_{v\leftrightarrow a} \quad (6)$$

To verify the effectiveness of the tri-tower contrastive pre-training framework, we also compare it with CLAP on two related cross-modal retrieval tasks: text-vocal retrieval and text-accompaniment retrieval. We report the experimental results in section 5.4.3, which indicates that including both vocal and accompaniment helps the model learn to ground more attribute-related song concepts.

5 Experiments

5.1 Dataset

To our knowledge, there are no public datasets available for controllable text-to-song. We crawl five thousand Mandarin songs covering around fifty singers, their lyrics, and some attribute tags (mood, instruments, style, etc.) from a well-known music website. There are 180 hours of audio data in total. In order to get the desired input, we perform some filtering and processing operations on the data. We present the details of data analysis and processing in Appendix B.2.

To alleviate data scarcity, we also leverage some open-source Mandarin singing voice datasets, which are listed in Appendix B.1.

5.2 Training and Evaluation

Model Configurations. For the tri-tower contrastive learning framework, we adopt the base version of BERT(Devlin et al., 2018) as the text encoder and the modified version of Audio Spectrogram Transformer(Gong et al., 2021) as the architecture of vocal encoder and accompaniment encoder. The [CLS] token from the final layer is projected into the joint embedding space of size 128. SoundStream(Zeghidour et al., 2021) has 12 quantization levels, each with a codebook of 1024 entries. The first three quantization levels are employed as acoustic tokens. The unit-based vocoder is built from the modified V1 version of BigVGAN(Lee et al., 2022). A comprehensive illustration of model hyperparameters is available in Appendix A.1.

Experimental Setup. We apply Spectrogram augmentation and text augmentation strategies for better performance. It takes 30 epochs for tri-tower pre-training using 8 NVIDIA V100 GPUs with a batch size of 128. For the training of text-to-song synthesis, we train the SVS model for 80K steps and the vocal-to-accompaniment model for 60K steps, both using 6 NVIDIA V100 GPUs with a batch size of 5000 tokens for each GPU. Each unit-based vocoder is trained using 4 NVIDIA V100 GPUs for 150K steps until convergence. The detailed setup is presented in Appendix A.2.

Evaluation. We conduct both subjective and objective evaluations on generated samples.

For the subjective evaluation, we conduct a crowd-sourced human evaluation via Amazon Mechanical Turk. We adopt the metrics of Mean Opinion Score (MOS) and Similarity Mean Opinion Score (SMOS) both with 95 % confidence intervals to evaluate the generated singing voice, which measures sample quality and speaker similarity respectively. Regarding the evaluation of accompaniment synthesis, we asked the rates to evaluate the audio samples in terms of overall quality (OVL), relevance to the prompt (REL), and alignment with the melody (MEL.) of the singing voice. For the objective evaluation of generated audios, we calculate the F0 Frame Error (FFE), Fréchet Audio Distance (FAD), Kullback–Leibler Divergence (KLD), and the CLAP score (CLAP).

Model	MOS (↑)	SMOS (↑)	FFE (↓)
GT	4.02± 0.05	/	/
FFT-Singer	3.71± 0.08	3.79±0.07	0.20
DiffSinger	3.80± 0.06	3.85±0.08	0.18
VISinger	3.82± 0.05	3.86±0.05	0.15
Make-A-Voice	3.86± 0.04	**3.89±0.08**	0.11
Melodist	**3.90±0.06**	3.87±0.07	**0.09**

Table 1: Ojective and subjective evaluation for Melodist and SVS baselines.

5.3 Singing Voice Synthesis

We compare our SVS model with four recent SVS baselines: 1) FFT-Singer, which generates mel-spectrograms through stacked feed-forward transformer blocks; 2) DiffSinger(Liu et al., 2022), which was built on diffusion probabilistic models to generate mel-spectrograms; 3) VISinger(Zhang et al., 2022b), an end-to-end singing synthesis model 4) Make-A-Voice(Huang et al., 2023), a multimodal spoken large language model for synthesizing and manipulating voice signals. Note that we also train a BigvGAN vocoder on 16k audios for FFT-Singer and DiffSinger to reconstruct waveform from Mel-spectrograms.

As shown in Table 1, our SVS model outperforms other baseline models with the highest MOS score of 3.89, indicating that it enjoys great superiority in sample quality. The SMOS score lags behind that of Make-A-Voice by a narrow margin but is better than other baseline models. The highest FFE score demonstrates the proficiency of Melodist in emulating the pitch prompt.

5.4 Vocal-to-accompaniment Synthesis

We report the comparison with text-to-music baselines. We include comparisons of different pre-trained text encoders. Moreover, cross-modal re-

Model	FAD (↓)	KLD (↓)	CLAP (↑)	OVL. (↑)	REL. (↑)	MEL (↑)
MUSICGEN (T5)	4.28	1.48	0.27	81.12±1.34	83.06±1.70	67.72±1.23
MUSICGEN (CLAP)	4.97	1.61	0.33	78.64±1.02	85.01±1.43	61.29±0.83
Melodist (T5)	**3.69**	1.36	0.29	**83.87±1.23**	83.58±1.61	78.05±0.75
Melodist (CLAP)	4.10	1.59	0.34	78.75±1.54	85.19±1.23	70.33±0.92
Melodist (Tri-Tower)	3.80	**1.34**	**0.39**	83.15±1.46	**86.63±1.27**	**79.40±0.96**

Table 2: Objective and Subjective evaluation of accompaniment samples generated by Melodist and MUSICGEN.

trieval result is presented to verify the effectiveness of the tri-tower contrastive pre-training framework.

5.4.1 Comparison to baselines

To our knowledge, SingSong(Donahue et al., 2023) is the only model with the same experimental setup as ours. However, its code and dataset are not available. So we only compare our model with MUSICGEN(Copet et al., 2023), the state-of-the-art single-stage controllable music generation model that can be conditioned on text and melody. Specifically, we adopt the vocal track extracted by Demucs as the melody condition of MUSICGEN. Following (Copet et al., 2023), we also investigate the impact of the text encoder for a fair comparison. We report a comparison with three text encoders in Table 2: 1) T5(Raffel et al., 2020), which is a Transformer architecture using a text-to-text approach; and 2) CLAP(Elizalde et al., 2023), a model for learning audio concepts from natural language supervision.

In general, Melodist surpasses MUSICGEN in objective and subjective metrics when applying the same text encoder, indicating the superiority of flattening prediction compared to the codebook interleaving strategies proposed in MUSICGEN. It reaches a trade-off between performance and computational efficiency.

Melodist presents the highest perceptual quality with outperformed FAD and OVL evaluation. When equipped with the text encoder of the tri-tower framework, the FAD and OVL scores drop slightly but still present better performance compared to MUSICGEN.

The adherence to the prefix condition can be witnessed in the evaluation result. Regarding to text prompts, Melodist outperforms MUSICGEN with the highest CLAP and REL scores and the lowest KLD score. Regarding to melody evaluation, the experimental results suggest that Melodist scores the best alignment with the melody of input, indicating that it can successfully generate accompaniments in harmony with the singing voice in melody.

5.4.2 Comparison of Different Text Encoder

As presented in Table 2, there is a subtle gap between the FAD and OVL scores of T5 and the text encoder of Tri-tower framework. We argue that this may be attributed to the pre-training data and scheme. In terms of adherence to text prompts, the Tri-tower framework outperforms other text encoders with the highest CLAP and REL score and the lowest KLD score. The superiority of the Tri-tower framework can be witnessed. It indicates that Melodist is capable of generating accompaniments that share similar semantic concepts with the text prompts while ensuring favorable audio quality. It is interesting to see that the text encoders trained in the contrastive learning paradigm show a better alignment between generated audios and text prompts, which demonstrates that the contrastive pre-training scheme significantly enhances text-guided music generation.

5.4.3 Cross-modal Retrieval Result

To further verify the effectiveness of the tri-tower contrastive framework, We conduct experiments of text-vocal retrieval and text-accompaniment retrieval. Specifically, we use 1K recordings as the pool of candidates and the paired vocal or accompaniment as the ground truth. We compare our tri-tower contrastive framework with three baselines: 1) MusCALL(Manco et al., 2022), a contrastive audio-language framework for Music; 2) MULAN(Huang et al., 2022a), a music audio and natural language joint embedding model; 3) CLAP(Elizalde et al., 2023), a model for learning audio concepts from natural language supervision. The sentence-level retrieval performance is evaluated by: 1) measuring mean average precision (mAP) for accuracy evaluation; and 2) Recall at the top k ranks (Recall@k). We set k to 1, 5, and 10.

As presented in Table 4, a significant superiority can be observed from these recall rates and the

Model of Stage 2	FAD (↓)	KLD (↓)	CLAP (↑)	OVL. (↑)	REL. (↑)	MEL (↑)
MUSICGEN	3.97	1.39	0.27	82.33±1.05	82.92±1.45	65.08±0.74
Melodist	**3.81**	**1.34**	**0.39**	**84.28±1.70**	**85.72±1.29**	**75.86±1.06**

Table 3: Objective and Subjective evaluation of song samples generated by Melodist and MUSICGEN.

mean average precision, indicating that including both vocal and accompaniment helps the model learn to ground more attribute-related song concepts. Jointly learning from vocals and accompaniments facilitates the text encoder extracting more accurate text representations of global characteristics, which greatly assists in subsequent vocal-to-accompaniment modeling. In addition, it is interesting that better retrieval performance is presented in text-to-accompaniment retrieval. This is mainly due to the reason that the text descriptions are more relevant to the accompaniment.

5.5 Text-to-song Synthesis

After a stage-by-stage evaluation, we compare the songs generated by Melodist and MUSICGEN in general terms. We fix the singing voice synthesis stage and generate the accompaniments with MUSICGEN and Melodist respectively. The only difference lies in the vocal-to-accompaniment model used for vocal-to-accompaniment synthesis. As we can see in Table 3, Melodist presents the highest perceptual quality and the best adherence to text prompt. It is identical to the observation of the previous section that Melodist outperforms MUSICGEN with outperformed scores, which is identical to the observation of the previous section.

Model	Recall (↑)			mAP (↑)
	@1	@5	@10	
Text-to-vocal Retrieval				
MusCALL	6.5	20.6	31.3	12.2
MULAN	8.2	22.7	34.5	13.0
CLAP	5.4	17.9	29.6	9.8
Melodist	**9.8**	**25.1**	**40.4**	**16.3**
Text-to-accompaniment Retrieval				
MusCALL	7.4	23.1	36.0	13.9
MULAN	8.0	22.3	38.2	15.3
CLAP	6.8	21.5	36.9	13.0
Melodist	**11.2**	**28.0**	**43.9**	**19.4**

Table 4: The experimental results of text-vocal retrieval and text-accompaniment retrieval.

5.6 Ablation

In this section, we investigate the impact of different data combinations and different augmentation strategies. Details and experimental results of the ablation can be found in the Appendix D.

Data Combination. We consider four combinations of crawled data and open-source data. We found that the absence of open-source SVS data leads to worse SVS performance, while a noticeable performance degradation in terms of audio quality and adherence can be witnessed when excluding open-resource song data.

Data Augmentation Strategies. We explore the effectiveness of text augmentation and spectrogram augmentation. When analyzing the experimental results, we can see a decline in both recall and mAP scores. A noticeable gain can be witnessed when applying data augmentation strategies.

6 Conclusion

In this paper, we introduce a new task called *text-to-song*, which incorporates singing voice and accompaniment synthesis from music score. We propose Melodist, the first text-to-song model with a two-stage generation scheme. Natural language prompts serve as the condition to control accompaniment generation. Melodist leverage a tri-tower contrastive pre-training framework to align the attribute prompt with its vocal and accompaniment in terms of the global pattern. We build up a Mandarin song dataset from the music website and leverage some open-source song and singing datasets to alleviate the data scarcity. We have conducted a series of comprehensive evaluations and the results indicate that Melodist outperforms baselines with comparable audio quality, temporal correspondence, and consistency with text concept. We provide extensive experiments to demonstrate the effectiveness of the tri-tower contrastive learning framework as well as the impact of different data combination and data augment strategies. In the future, we will focus on improving the audio quality and vocal accompaniment harmonization.

Limitations

Though Melodist have shown comparable achievements in text-to-song synthesis, its limitations cannot be ignored. The reliance on source separation imposes a great challenge to improving audio quality. While the current source separation methods remain suboptimal, it is urgent to improve the quality of source separation. There are some alternatives such as constructing a high-quality dataset or designing a fully end-to-end text-to-song synthesis model. Additionally, Melodist treats accompaniment as a single track, disregarding the intricate composition of individual elements such as drums, bass, and other instrument-related tracks. A promising avenue for future exploration involves both intra-track and inter-track modeling, thereby facilitating a more comprehensive approach to text-to-song synthesis.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

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.

Tianyu Chen, Yuan Xie, Shuai Zhang, Shaohan Huang, Haoyi Zhou, and Jianxin Li. 2022. Learning music sequence representation from text supervision. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4583–4587. IEEE.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*.

Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. 2019. Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254*.

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*.

Fanyu Ding and Yidong Cui. 2023. Museflow: music accompaniment generation based on flow. *Applied Intelligence*, 53(20):23029–23038.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023. Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*.

Chris Donahue, Antoine Caillon, Adam Roberts, Ethan Manilow, Philippe Esling, Andrea Agostinelli, Mauro Verzetti, Ian Simon, Olivier Pietquin, Neil Zeghidour, et al. 2023. Singsong: Generating musical accompaniments from singing. *arXiv preprint arXiv:2301.12662*.

Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. 2018. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. 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.

Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM*, 63(11):139–144.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE.

Zhiqing Hong, Chenye Cui, Rongjie Huang, Lichao Zhang, Jinglin Liu, Jinzheng He, and Zhou Zhao. 2023. Unisinger: Unified end-to-end singing voice synthesis with cross-modality information matching. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7569–7579.

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. 2022a. Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415*.

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*, MM '21, page 3945–3954, New York, NY, USA. Association for Computing Machinery.

Rongjie Huang, Chenye Cui, Feiyang Chen, Yi Ren, Jinglin Liu, Zhou Zhao, Baoxing Huai, and Zhefeng Wang. 2022b. Singgan: Generative adversarial network for high-fidelity singing voice generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2525–2535.

Rongjie Huang, Chunlei Zhang, Yongqi Wang, Dongchao Yang, Luping Liu, Zhenhui Ye, Ziyue Jiang, Chao Weng, Zhou Zhao, and Dong Yu. 2023. Make-a-voice: Unified voice synthesis with discrete representation. *arXiv preprint arXiv:2305.19269*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033.

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*.

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.

Gurunath Reddy Madhumani, Yi Yu, Florian Harscoët, Simon Canales, and Suhua Tang. 2020. Automatic neural lyrics and melody composition. *arXiv preprint arXiv:2011.06380*.

Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. 2022. Contrastive audio-language learning for music. *arXiv preprint arXiv:2208.12208*.

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.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pages 5210–5219. PMLR.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Yi Ren, Jinzheng He, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. Popmag: Pop music accompaniment generation. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1198–1206.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2020. AISHELL-3: A multi-speaker mandarin TTS corpus and the baselines. *CoRR*, abs/2010.11567.

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*.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. 2022. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE.

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

Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. 2017. Midinet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv preprint arXiv:1703.10847*.

Lili Yu, Dániel Simig, Colin Flaherty, Armen Aghajanyan, Luke Zettlemoyer, and Mike Lewis. 2023. Megabyte: Predicting million-byte sequences with multiscale transformers. *arXiv preprint arXiv:2305.07185*.

Yi Yu, Abhishek Srivastava, and Simon Canales. 2021. Conditional lstm-gan for melody generation from lyrics. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(1):1–20.

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.

A The Details of Experiment

A.1 Model Configuration

The model hyper-parameters of Melodist are listed in Table 5.

A.2 Experimental Setup

In Tri-tower contrastive pretraining, each audio is converted to a log-scaled mel spectrogram with the FFT size of 1024, hop size of 256, and window size of 1024. We then chunk the augmented spectrogram into 16× 16 patches. We limit the max text sequence length to 77 chars for computational efficiency. Inspired by (Copet et al., 2023), text augmentation is applied by concatenating tag lists to the text description. We limit the max text sequence length to 77 chars for computational efficiency. A [CLS] token is prepended to the sequence as a summary of the contextual patch embeddings in three encoders. We set the temperature τ to 0.2. For two-stage text-to-song synthesis, the learning rate is set to 5e-5. Adam optimizer is used with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$.

The unit-based vocoder is trained on 16k audio data with a segment size of 32000. The learning rate is set to 5e-5. Adam optimizer is used with $\beta_1 = 0.8$, $\beta_2 = 0.99$, and $\epsilon = 10^{-6}$.

B Dataset Analysis

In this section, we describe the details of the dataset for training.

B.1 Open-Source Datasets

We present the open-source datasets adopted for training in Table 6.

B.2 The Crawled Song Data

B.2.1 Data processing pipeline

In order to get the desired input, we perform the following filtering and processing operations on the data:

Data Filtering. We exclude audios that 1) are live songs; 2) of silent accompaniment or no vocals; 3) are performed by multiple singers. Additionally, some content (composer, performer, etc.) irrelevant to text transcriptions is removed from the lyrics.

Source Separation. We split each song into 10-second clips from each song and passed each clip to the Demucs(Défossez et al., 2019) to separate vocals from the rest of the accompaniments and yield aligned pairs of waveforms. Finally, we resample vocal and instrumental clips from 44.1kHz to 16kHz and average all audio files to mono.

Lyrics-to-Singing Alignment. We first reorganize the clips of the separated vocals and restore them to the original songs. Then we use Montreal forced alignment(McAuliffe et al., 2017) tool to extract the phoneme duration. After filtering the misaligned segments, we segment each song in 6-10s according to the separation marks in raw lyrics.

Prompt Generation. We copy the tags of a song to its segments and then make minor modifications according to the auditory impression. A tag-to-pseudo caption generation approach with large language models(Doh et al., 2023) is leveraged to generate natural language prompts.

B.2.2 Examples of Prompt

We provide some examples of attribute tag lists and the captions generated by (Doh et al., 2023).

There are examples of crawled attribute tag lists:

- pop, bass, guitar, acoustic, beat.
- rock, passionate, vocal, shimmering, bass, guitar, acoustic, guitar, guitar, emotional, passionate.
- instrumental, melodic, saxophone, acoustic, guitar, soft, mellow, ambient, dreamy.
- cool, vocal, bass, percussion, retro, dance.
- guitar, synth, bass, guitar, electronic, beat, sentimental, dance, club

There are examples of generated text descriptions:

- This is a pop music piece. There is a male vocalist singing melodically in the lead. The melody is being played by the keyboard while the bass guitar is playing in the background. The rhythm consists of a slow tempo electronic drum beat. The atmosphere is easygoing. This piece could be used in the soundtrack of a romantic comedy movie, especially during the scenes where a character is hesitating to open up to their crush.
- The low quality recording features a rock song that consists of a passionatele vocal singing over punchy kick and snare hits, shimmering hi hats, soft kick and groovy bass guitar. It sounds addictive, energetic and passionate.
- This music is a Jazz instrumental. The tempo is slow with a melodic saxophone harmony, keyboard accompaniment and rhythmic acoustic guitar accompaniment. The music is soft, mellow, pleasant, ambient, dreamy and pleasant.

Hyperparameter		Melodist	Number of parameters
Global Transformer	Hidden Size Layers Hidden Dim Attention Heads FFN Dim	192 20 1152 16 4608	320.07M
Local Transformer	Hidden Size Layers Hidden Dim Attention Heads FFN Dim	192 6 1152 8 4608	100.14M
Unit-based Vocoder	Upsample Rates Hop Size Upsample Kernel Sizes	[5, 4, 2, 2, 2, 2] 320 [9, 8, 4, 4, 4, 4]	121.60M
Vocal Encoder	Layers Hidden Dim Attention Heads FFN Dim	6 768 8 3072	42.10M

Table 5: Hyperparameters of Melodist.

Dataset	Type	Annotation	Volume (hrs)
Stage 1: Singing Voice Synthesis			
Opencpop (Wang et al., 2022)	singing	text, duration, MIDI	5.2
M4Singer (Zhang et al., 2022a)	singing	text, duration, MIDI	29.8
OpenSinger (Huang et al., 2021)	singing	text, duration, MIDI	86.5
PopCS (Liu et al., 2022)	singing	text, duration	5.9
AISHEELL-3 (Shi et al., 2020)	speech	text	85
Stage 2: Vocal-to-accompaniment Synthesis			
LP-MusicCaps-MSD (Doh et al., 2023)	music	text description	7k

Table 6: Statistics of training datasets.

- A female singer sings this cool melody with backup singers in vocal harmony. The song is medium tempo with a steady drumming rhythm, keyboard accompaniment, percussive bass line and various percussion hits. The track is a retro hip hop dance tune.
- This is an amateur recording of a R&B music piece. There is a male vocalist singing melodically in the lead. The melody is being played by the electric guitar and the synth bass guitar while the rhythmic background consists of a slow tempo electronic drum beat. The atmosphere is sentimental. This piece could be playing in the background at a nightclub or a dance club.

C Evaluation

C.1 Subjective Evaluation

We randomly selected 30 audio samples generated from each stage and each sample was evaluated by 20 raters via Amazon Mechanical Turk. We paid $8 an hour for participant compensation.

For quality evaluation of generated singing voice, we conduct the MOS (mean opinion score) tests and explicitly instruct the raters to "(focus on examining the audio quality and naturalness, and ignore the differences of style (timbre, emotion, and prosody).)". The testers present and rate the samples, and each tester is asked to evaluate the subjective naturalness on a 1-5 Likert scale.

For speaker similarity evaluation, we ask the raters to focus on the similarity of the speaker identity (timbre) to the reference and ignore the

ID	SVS Data	Song Data	Stage 1		Stage 2			
			MOS (↑)	FFE (↓)	FAD (↓)	KLD (↓)	OVL (↑)	REL (↑)
1	✓	✗	3.89±0.08	0.09	3.88	1.46	79.56±1.42	83.02±1.39
2	✗	✓	3.84±0.05	0.13	3.79	1.39	83.10±1.31	86.56±1.80
3	✗	✗	3.84±0.05	0.13	3.88	1.46	79.56±1.42	83.02±1.39
4	✓	✓	**3.89±0.08**	**0.09**	**3.79**	**1.39**	**83.10±1.31**	**86.56±1.80**

Table 7: Ablation study on different data combination.

differences in content, grammar, or audio quality. We paired each synthesized utterance with a reference utterance to evaluate how well the synthesized speech matched that of the target speaker.

For the evaluation of generated accompaniments, we follow (Copet et al., 2023) to evaluate overall quality (OVL), and relevance to the text input (REL). In terms of alignment with the melody (MEL.), we ask the rates to focus more on temporal correspondence between accompaniment and reference singing voice instead of melody resemblance.

The Screenshot of subjective evaluation is presented in Figure 4, 5. A small subset of samples used in the test is available at https://research.github.io/text-to-song/.

D Ablation Study

Data Combinations. We consider four combinations of crawled data and open-source data. when training Melodist, including 1) Exclude open-source SVS data in Stage 1; 2) Exclude song data in Stage 2; 3) Exclude open-source SVS and song data; 4) Include open-source SVS and song data as the original setting.

Data Augmentation. We explore the effectiveness of text augmentation and spectrogram augmentation.

We report the evaluation results in Table 7 and Table 8. It suggests that leveraging open-source datasets and augmentation strategies enhance the capability of Melodist to generate more high-fidelity and consistent output.

Model	Recall (↑)			mAP (↑)
	@1	@5	@10	
Text-to-vocal Retrieval				
w/o TA	6.7	18.2	34.2	13.7
w/o SA	8.0	20.6	33.9	12.2
w/o TA&SA	6.3	15.8	32.3	10.3
TA&SA	**9.8**	**23.7**	**40.2**	**15.7**
Text-to-accompaniment Retrieval				
w/o TA	7.4	21.1	37.0	14.5
w/o SA	8.5	22.3	39.1	15.9
w/o TA&SA	6.2	18.5	35.9	13.1
TA&SA	**11.3**	**27.6**	**41.1**	**19.4**

Table 8: Ablation study on the impact of data augmentation strategies. We report the experimental results of text-vocal retrieval and text-accompaniment retrieval. SA denotes spectrogram augmentation and TA denotes text augmentation.

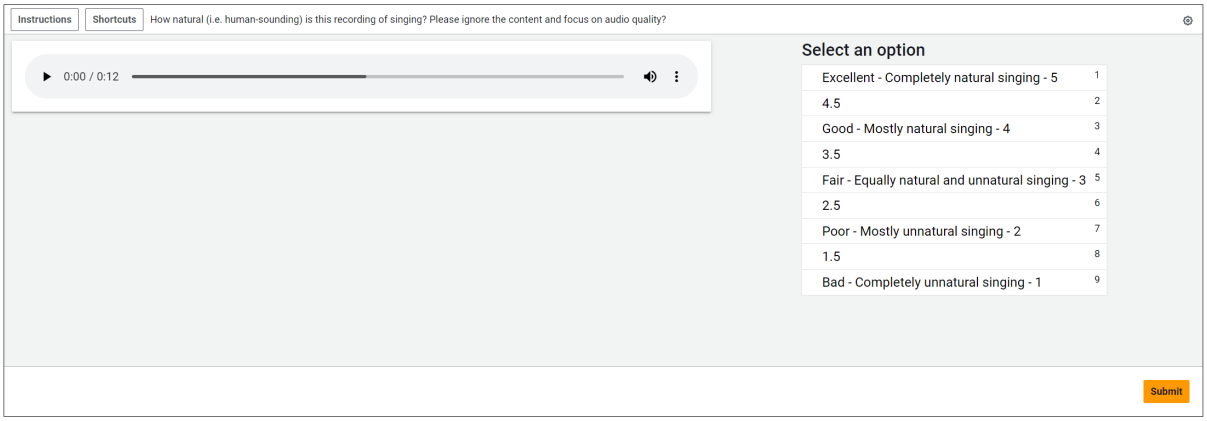

Figure 4: Screenshot of MOS testing.

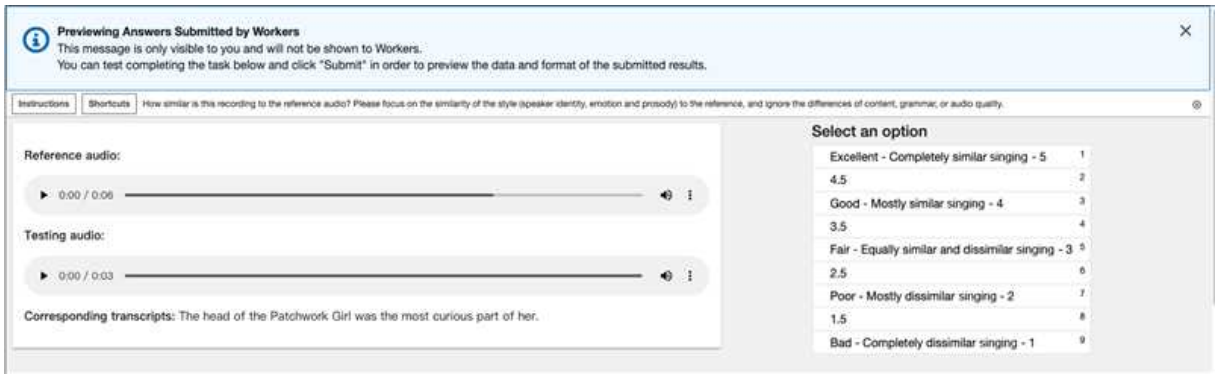

Figure 5: Screenshot of SMOS testing.