SingingSDS: A Singing-Capable Spoken Dialogue System for Conversational Roleplay Applications

Jionghao Han[1], Jiatong Shi[1], Masao Someki[1], Yuxun Tang[2], Lan Liu[2], Yiwen Zhao[1], Wenhao Feng[2], Shinji Watanabe[1],

[1]Carnegie Mellon University, [2]Renmin University of China

Editors: D. Herremans, K. Bhandari, A. Roy, S. Colton, M. Barthet

Abstract

With recent advances in automatic speech recognition (ASR), large language models (LLMs), and text-to-speech (TTS) technologies, spoken dialogue systems (SDS) have become widely accessible. However, most existing SDS are limited to conventional spoken responses. We present **SingingSDS**, a cascaded SDS that responds through singing rather than speaking, fostering more affective, memorable, and pleasurable interactions in character-based roleplay and interactive entertainment scenarios. SingingSDS employs a modular ASR-LLM-SVS pipeline and supports a wide range of configurations across character personas, ASR and LLM backends, SVS models, melody sources, and voice profiles, tailored to different needs in terms of latency, quality, and musical style. SingingSDS is available as a plug-and-play web demo, featuring modular, open-source code that supports customization and extension. Demo: https://huggingface.co/spaces/espnet/SingingSDS. Code: https://github.com/SingingSDS/SingingSDS. Video: https://youtube.com/playlist?list=PLZpUJJbwp2WvtPBenG5D3h09qKIrt24ui&si=7CSLWAYWcfkTEdqe.

Keywords: Spoken Dialogue System, Singing Voice Synthesis, Large Language Models, Speech-to-Singing, Interactive Roleplay

1. Introduction

Spoken dialogue systems (SDS) have seen rapid advancements in recent years (Yu et al., 2025; Ding et al., 2025; Xu et al., 2025; Arora et al., 2025; Li et al., 2025; Gao et al., 2025; Défossez et al., 2024), with increasing focus on role-play, character embodiment, and immersive interaction (Huang et al., 2025; Zhang et al., 2025; Chiang et al., 2025). Such systems have demonstrated their potential in enhancing user engagement through dynamic and emotionally expressive conversations, as exemplified by character-driven applications like Neuro-sama (Neuro-sama, 2024), Cotomo (Starley Co.), and interactive experiences such as Turtle Talk with Crush (Walt Disney World Resort). However, conventional SDS outputs are typically limited to standard speech, which constrains the potential for richer, aesthetically engaging experiences.

Singing, as a communicative modality, combines linguistic content with melody and rhythm, offering enhanced memorability and pleasure compared to speech (Haiduk et al., 2020; Gold et al., 2019; Zatorre and Salimpoor, 2013), which can enrich interactive entertainment experiences. Despite significant progress in singing voice synthesis (SVS) and song generation models (Yuan et al., 2025; Wu et al., 2024a,b; Tang et al., 2024b; Yu et al.,

2024), these systems are essentially non-interactive: largely operate on predefined lyrics, lacking mechanisms for dynamic responses to user input.

To address this gap, we introduce **SingingSDS**, the first open-source system supporting speech-in, singing-out roleplay interactions for entertainment and character-driven scenarios. SingingSDS integrates automatic speech recognition (ASR), character-consistent response generation using large language models (LLMs), melody control with optional structural constraints, and singing voice synthesis (SVS). The system is modular and configurable, including 5 ASR models, 7 LLMs, our released bilingual (Chinese-Japanese) and monolingual (Chinese-only) SVS models, and 5 melody control settings, resulting in 350 possible system configurations. We conduct systematic assessment of both audio quality and user perception, supporting reproducible research on interactive singing dialogue. The system is fully open-sourced and provides an interactive web demo and a command-line interface for the creation and evaluation of speech-to-singing dialogues with fictional characters. These features support reproducible research and structured experimentation with interactive singing dialogues.

SingingSDS establishes a foundation for investigating singing as an interactive response modality beyond conventional spoken dialogue. The system has potential applications in VR concerts and other virtual performances, interactive music games and theme park attractions, and live streaming with audience participation. Through singing responses, SingingSDS can enhance these applications, offering more memorable and enjoyable user experiences, while also providing a platform for empirical studies of singing-based dialogue.

2. Related Work

Conventional SDS have been widely adopted in AI-assisted applications (Apple Inc., 2025; Amazon.com., 2025; Google, 2025). Recent advancements in SDS (Yu et al., 2025; Ding et al., 2025; Xu et al., 2025; Arora et al., 2025; Li et al., 2025; Gao et al., 2025; Huang et al., 2025; Zhang et al., 2025; Chiang et al., 2025) have improved these systems' fluency and coherence, but they largely remain focused on usual conversational interactions, with limited exploration of creative modalities such as singing.

In parallel, SVS has progressed significantly in recent years with the development of neural models such as TokSing (Wu et al., 2024b), DiffSinger (Liu et al., 2022), and VISinger2 (Zhang et al., 2022b), which enable high-fidelity singing generation by modeling pitch, duration, and timbre. Despite the progress in both SDS and SVS, to the best of our knowledge, no prior work has integrated singing voice synthesis into an interactive spoken dialogue system. Our work presents the first attempt to bridge these two domains, enabling an LLM-based dialogue agent to sing its responses to the user via SVS techniques.

One of the key challenges in equipping LLM-based spoken dialogue systems with singing capabilities lies in evaluation. While various metrics have been proposed to assess synthesized speech and singing quality (Saeki et al., 2022; Umbert et al., 2015; Tang et al., 2024a; Shi et al., 2025, 2024b), existing tools often fail to account for the entertainment value conveyed through singing or speech.

In our experiments, model-based metrics such as Meta AudioBox Aesthetics (Tjandra et al., 2025) did not consistently align with human preferences, and in some cases favored randomly generated, inharmonic note sequences over well-structured melodies. To better

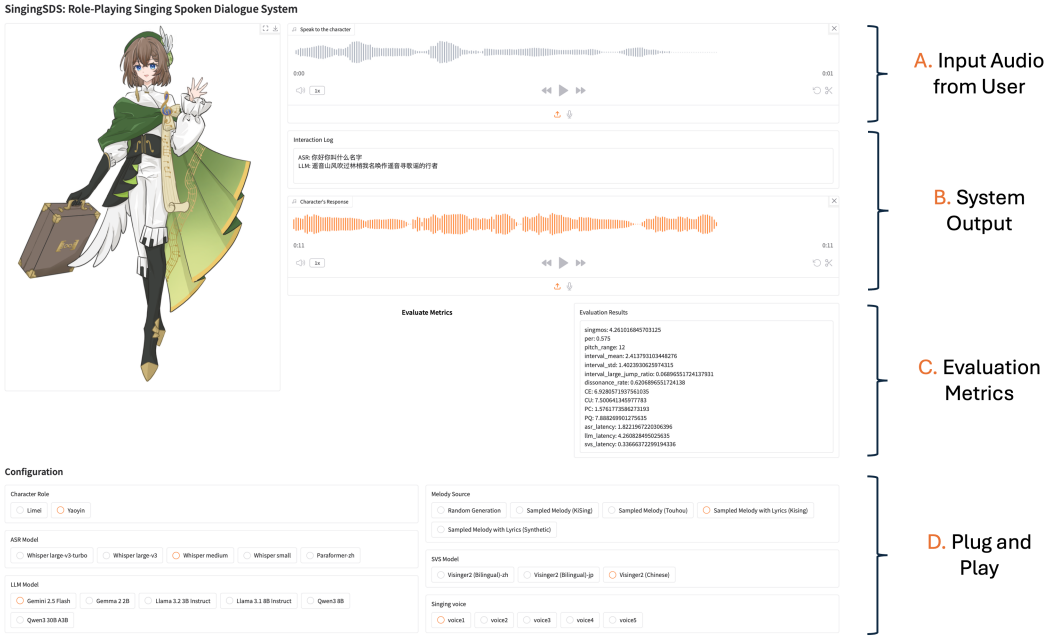


Figure 1: Web interface of SingingSDS. (A) User audio input via microphone or file upload. (B) Visualization of ASR transcription, LLM-generated response, and SVS-generated singing. (C) Evaluation results. (D) Configurable interface for selecting characters, models, voices, and melodies.

capture the aspects of engagement and enjoyability, we conducted human evaluations focusing on perceived enjoyment. Additionally, we report coarse melodic statistics, such as the large jump ratio, to quantify pitch dynamics in the generated singing outputs. Together, these complementary metrics offer a more holistic perspective on melody-conditioned dialogue generation (Appendix E).

3. System Design

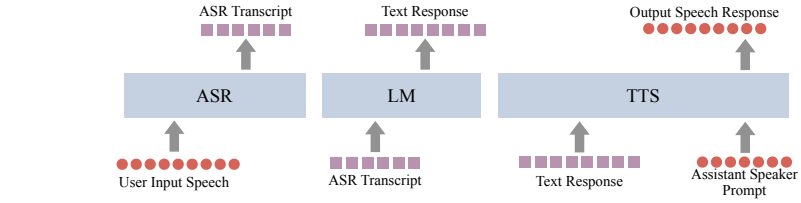
Based on the requirements, our system adopts a cascaded ASR-LLM-SVS pipeline with reference melodies (Figure 2). Additional architectural considerations and design trade-offs are discussed in Appendix A.

ASR. Given a user speech input s and the specified language ℓ, the system first transcribes the utterance using an ASR module:

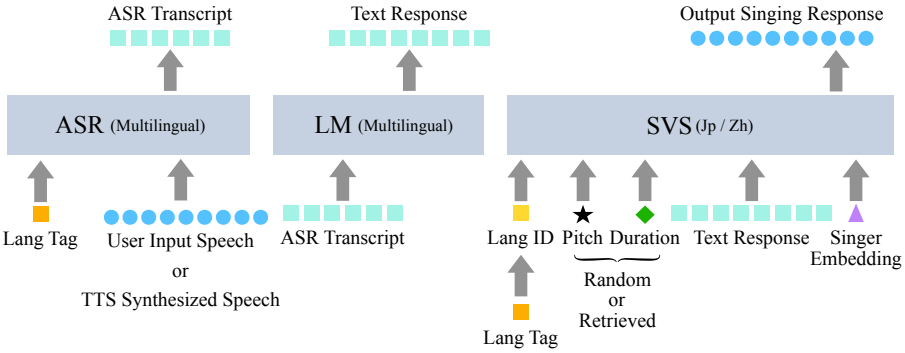
$$s_t = \mathrm{ASR}(s, \ell)$$

ℓ is explicitly provided to avoid errors from language identification capabilities of an ASR backend and improves recognition accuracy within a dialogue.

LLM. The transcription s_t is then passed to a LLM, which generates an in-character reply conditioned on the user's utterance, the virtual character's profile c, and optional structural



(*a*) Baseline SDS (Arora et al., 2025)



(*b*) Our singing dialogue system

Figure 2: Comparison of (a) the baseline spoken dialogue system with (b) our proposed extension to support singing dialogue.

constraints $\mathcal{C}$. The model is prompted with a system prompt containing $(c, \mathcal{C})$ and a user prompt containing s_t:

$$l = \text{LLM}(\texttt{SystemPrompt}(c, \mathcal{C}),\ \texttt{UserPrompt}(s_t))$$

The character's profile c largely follows the standard persona format used in OmniCharacter (Zhang et al., 2025), with adaptations tailored to lyrical dialogue in our system. The structural constraint $\mathcal{C}$ is derived from the melody controller when phrase annotations are available. Full prompt templates are provided in Appendix B.

Melody Control. The melody controller provides note-level constraints in the form of a sequence $\mathcal{N} = (p_i, \tau_i^s, \tau_i^e)_{i=1}^n$, where p_i denotes pitch, and τ_i^s, τ_i^e indicate the start and end times (in seconds) of each note. Optional phrase annotations define the boundaries of musical phrases.

We support two types of melody sources. The first setting consists of randomly synthesized melodies and serves as a baseline. These are generated on the fly by sampling pitch and duration values uniformly, without rests or phrase-level structure. Since no reference alignment is available, a simple forced alignment is applied, assigning one syllable per note. The second is sampled melodies drawn from existing song datasets. For these, we support two alignment strategies. In pitch-based alignment, each syllable is mapped one-to-one to a note in the melody. In lyric-aware alignment, one-to-many mappings are preserved: when a syllable spans multiple notes in the original song, the same structure is retained in the output, as illustrated in Figure 3.

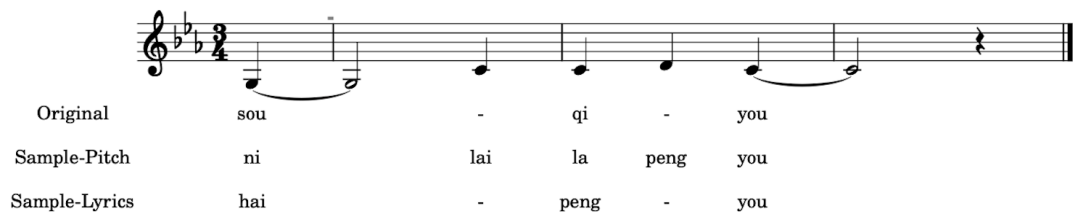


Figure 3: Illustration of melody alignment strategies on a two-bar phrase from the KiSing dataset. The melody is shared across all rows. The top row shows the original lyrics from the dataset. The middle row displays the alignment under `sample-pitch`, where each syllable is mapped one-to-one to a note. The bottom row corresponds to `sample-lyric`, where original multi-note syllables are preserved to match the source phrasing. A dash "–" indicates that the preceding syllable is sustained over the current note (i.e., extended phonation).

To encourage structural alignment between the generated textual response and the melody, the system constructs an LLM prompt specifying the required syllable count per musical phrase when phrase annotations are available (e.g., in the KiSing dataset and our self-constructed melody datasets synthesized with Yue (Yuan et al., 2025)). This alignment serves as a soft constraint for the LLM, encouraging outputs that match the expected number of musical events and exhibit more coherent phrase-level structure. Details on the phrase-constrained prompt used for LLM generation are provided in Appendix B.2.

SVS. The generated lyrical response l is normalized and converted into phonemes l_ϕ with grapheme-to-phoneme (G2P) system. Along with a music score $\mathcal{N}$ created by the melody controller module and speaker information v, either speaker embodding or speaker identity depending on the model, the inputs are passed to an SVS model, producing the final sung output:

$$\hat{S} = \text{SVS}(\text{MelodyControl}(l_\phi, \mathcal{N}), v)$$

4. Demonstration

SingingSDS adopts a modular architecture with registry-based components that enable flexible integration of models, datasets, and character personas. As shown in Figure 4, each core function, such as ASR, LLM, SVS, and melody loading and handling, is encapsulated as an independent module. This design supports rapid iteration, systematic benchmarking, and seamless extensibility.

4.1. Models

Our system supports multiple backends for ASR, LLM, and SVS, all integrated through a registry-based modular architecture. Most ASR and LLM modules are community-pretrained. The supported SVS models are trained by us.

```
SingingSDS/
├── app.py, cli.py              # Entry points (demo app & CLI)
├── pipeline.py                 # Main orchestration pipeline
├── interface.py                # Gradio interface
├── characters/                 # Virtual character definitions
├── modules/                    # Core modules
│   ├── asr/                    # ASR models (Whisper, Paraformer)
│   ├── llm/                    # LLMs (Gemini, LLaMA, etc.)
│   ├── svs/                    # Singing voice synthesis (ESPnet)
│   └── utils/                  # G2P, text normalization, resources
├── config/                     # YAML configuration files
├── data/                       # Dataset metadata and length info
├── data_handlers/              # Parsers for KiSing, Touhou, etc.
├── evaluation/                 # Evaluation metrics
├── resources/                  # Singer embeddings, phoneme dicts, MIDI
├── assets/                     # Character visuals
├── tests/                      # Unit tests and sample audios
└── README.md, requirements.txt
```

Figure 4: Modular architecture of our system. Each module (ASR, LLM, SVS, melody, character behavior) is encapsulated as a standalone component and connected through a central interface. A Gradio-based UI and YAML configuration templates facilitate rapid deployment and customization.

Table 1: Supported backend models in our system. ASR and dialogue components use publicly available pretrained models. SVS models were trained and released by us on Hugging Face.

Component	**Model Name**	**Source**
ASR	Whisper (Radford et al., 2023) (small, medium, large-v3, large-v3-turbo)	OpenAI
	Paraformer (Gao et al., 2022, 2023)	Alibaba
LLM	Gemini 2.5 Flash (DeepMind, 2025), Gemma 2 2B (Team et al., 2024)	Google
	Llama 3.2 3B Instruct, Llama 3.1 8B Instruct (Grattafiori et al., 2024)	Meta
	Qwen3 8B, Qwen3 30B A3B (Yang et al., 2025)	Alibaba
	MiniMax-Text-01 (MiniMax et al., 2025)	MiniMaxAI
SVS	VISinger 2 (CN, multi-speaker)	Ours (Hugging Face)
	VISinger 2 (CN/JP, multi-speaker)	Ours (Hugging Face)

We provide two multi-speaker VISinger 2 models (Zhang et al., 2022b): (1) a Chinese SVS model[1] trained on the ACE-Opencpop dataset (Shi et al., 2024a), and (2) a bilingual Mandarin-Japanese SVS model[2] trained on a mixture of publicly available singing datasets, including OpenCpop (Huang et al., 2021), KiSing (Shi et al., 2022), ACE-KiSing (Shi et al., 2024a), M4Singer (Zhang et al., 2022a), Kiritan (Ogawa and Morise, 2021), Onikuru Kurumi (Kurumi, 2020), PJS (Koguchi et al., 2020), and Namine Ritsu (Canon, 2009). Details of the training configuration are provided in the Appendix C.

A full list of supported models is summarized in Table 1.

1. https://huggingface.co/espnet/aceopencpop_svs_visinger2_40singer_pretrain
2. https://huggingface.co/espnet/visinger2-zh-jp-multisinger-svs

4.2. Datasets

The system supports retrieval from three melody dataset in addition to randomly generated melodies: KiSing (Shi et al., 2022), a Touhou MIDI collection[3], and a synthesized dataset of 499 songs generated using Yue, constructed to expand the melody database (see Appendix D for details). These datasets provide melodies that condition the singing output. A registry-based handler module loads and converts each melody into a format suitable for synthesis, allowing new datasets to be integrated with minimal effort.

4.3. Characters

Our system supports two original singing characters, Limei and Yaoyin, each defined by a prompt-based persona, as specified in Appendix B.1. Both characters are drawn from our original fictional universe, *Changge Plains*, designed to support immersive roleplay interaction and storytelling. New characters can be added by specifying prompt configurations.

4.4. Deployment and Access

SingingSDS is available as an interactive web demo hosted on Hugging Face Spaces.[4] Users can initiate dialogue by speaking into a microphone. The system transcribes the input, generates an in-character lyrical response, and synthesizes a singing reply. The interface displays synchronized lyrics, character portraits, and playback controls, shown in Figure 1. Users can switch between characters (e.g., Limei and Yaoyin) and different model and melody configurations.

In addition to the web demo, SingingSDS can be run locally in two modes:

- **Web app mode:** Install dependencies in `requirements.txt` and launch `app.py` for a local Gradio UI.
- **CLI mode:** Run `cli.py` for command-line usage. This supports non-interactive synthesis, dataset creation, and benchmarking.

All code and usage instructions are available at: https://github.com/SingingSDS/SingingSDS

Table 2: Evaluation results of different ASR (Whisper, Paraformer), LLM (Llama 3, Gemini) and melody (KiSing, Touhou) configurations. ↑ indicates higher is better; ↓ indicates lower is better. Note that the Large Jump Ratio (Jump R.) reflects melody dynamics and does not necessarily favor lower values. Detailed metric definitions can be found in Appendix E.

ASR	LLM	Melody	SingMOS ↑	PER (%) ↓	Jump R.	N&F ↑	Char. Cons. ↑	Lyric Qual. ↑	ASR Lat. ↓	LLM Lat. ↓	SVS Lat. ↓
Whisper	Llama 3	KiSing	4.53	0.61	0.11	4.00	4.17	3.35	0.80	1.87	0.19
Paraformer	Llama 3	KiSing	4.47	**0.12**	0.13	4.08	4.13	3.41	**0.44**	**1.79**	**0.16**
Whisper	Gemini	KiSing	**4.59**	0.48	0.09	**4.21**	**4.19**	**3.86**	0.55	5.79	0.18
Whisper	Llama 3	Touhou	4.52	0.14	0.28	4.06	4.13	3.70	0.82	2.22	0.19

3. https://github.com/AyHa1810/touhou-midi-collection
4. https://huggingface.co/spaces/espnet/SingingSDS

5. Evaluation

We evaluate the system from multiple perspectives, including perceptual quality, linguistic accuracy, melodic structure, and runtime efficiency through automated or human evaluation. Detailed explanations on evaluation setups can be found in Appendix E.

The evaluation module is fully integrated into the system and can be triggered directly through the user interface or computed with our CLI command.

5.1. Datasets

We evaluate SingingSDS on a self-constructed roleplay test set of 20 prompts, targeting our fictional persona Yaoyin to evaluate character-conditioned generation.

We also evaluated using a subset of KdConv dataset (Zhou et al., 2020), a multi-domain multi-turn dialogue corpus, to simulate user interactions. The experimental setup and results for KdConv sampled data can be found in Appendix F.

All audio outputs are resampled to 16 kHz for ASR-based intelligibility evaluation (i.e., PER) and kept at 44.1 kHz for subjective MOS testing. For melody selection, we use scores retrieved from the KiSing dataset and a curated archive of Touhou MIDI files.

5.2. Experimental Setup

Our experiments are run on single NVIDIA v100 GPU using the cascaded pipeline shown in Figure 2. We evaluate two ASR models: `whisper-medium` (OpenAI) and `paraformer-zh` (Alibaba), and two LLMs: `Llama-3.1-8B-Instruct` and `gemini-2.5-flash`. For brevity, we refer to them as Whisper, Paraformer, Llama 3, and Gemini in the rest of this paper. For singing voice synthesis (SVS), we use our bilingual pretrained VISinger 2 model.

5.3. Results and Discussion

We evaluate our system under multiple configurations of the ASR, LLM, and melody generation modules. Table 2 presents the performance across combinations of Whisper and Paraformer (ASR), LLaMA 3 and Gemini (LLM), and KiSing and Touhou (melody). The Whisper + Gemini configuration achieves the highest overall perceptual quality and entertainment value, as indicated by automatic singing quality scores (SingMOS) and human evaluations of novelty and fun (N&F), character consistency (Char. Cons.), and lyric quality (Lyric Qual.). In contrast, the Paraformer + LLaMA 3 setting yields the lowest system latency, making it more suitable for interactive scenarios.

6. Conclusion

This paper presents SingingSDS, a modular spoken dialogue system with melody-conditioned singing responses to user input in Chinese and Japanese. The system combines ASR, LLM, and SVS components through a prompting scheme that aligns lyric structure with melodic phrasing, without requiring model fine-tuning.

Evaluation across perceptual quality, intelligibility, latency, and melodic dynamics confirms the feasibility of singing-based interaction in dialogue systems. Subjective ratings

indicate that appropriate melody selection improves perceived entertainment value without compromising intelligibility.

To support future work, we release the pretrained SVS model used in SingingSDS, along with scripts for evaluation and dataset construction. Although other components are based on open-source APIs, the pipeline remains modular and extensible, allowing substitution of melody sources, LLMs, or SVS backends for controlled experimentation.

SingingSDS constitutes the first fully implemented pipeline for interactive dialogue system with singing virtual characters. By bridging conversational AI and singing synthesis, it enables a novel form of interactive response grounded in melody and persona. Our system opens new research directions in controllable singing generation, expressive speech interfaces, and musical human-computer interaction.

Acknowledgments

We acknowledge illustrator Zihe Zhou for the creation of Yaoyin's character artwork, which is included in the demo page shown in Figure 1. The artwork was commissioned exclusively for the SingingSDS project and may be used for direct derivatives of SingingSDS, such as project-related posts or usage videos, without additional permission. Any other use requires express permission from the illustrator. Use of the artwork for training or fine-tuning artificial intelligence or machine learning models is strictly prohibited.

Parts of the experiments of this work used the Bridges2 system at PSC through allocations CIS210014 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

Amazon.com. Alexa. https://www.amazon.com/dp/B0DCCNHWV5?ref=aucc_web_red_xaa_evgn_tx_0002, 2025. Accessed: 2025-06-26.

Apple Inc. Siri. https://www.apple.com/siri/, 2025. Accessed: 2025-06-26.

Siddhant Arora, Yifan Peng, Jiatong Shi, Jinchuan Tian, William Chen, Shikhar Bharadwaj, Hayato Futami, Yosuke Kashiwagi, Emiru Tsunoo, Shuichiro Shimizu, et al. Espnet-sds: Unified toolkit and demo for spoken dialogue systems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 248–259, 2025.

Canon. Namine ritsu singing voice database. https://drive.google.com/drive/folders/1XA2cm3UyRpAk_BJb1LTytOWrhjsZKbSN, 2009.

Cheng-Han Chiang, Xiaofei Wang, Chung-Ching Lin, Kevin Lin, Linjie Li, Radu Kopetz, Yao Qian, Zhendong Wang, Zhengyuan Yang, Hung-yi Lee, et al. Audio-aware large language models as judges for speaking styles. *arXiv preprint arXiv:2506.05984*, 2025.

Google DeepMind. Gemini 2.5 flash. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash, 2025. Accessed July 4, 2025.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. Technical report, 2024. URL https://arxiv.org/abs/2410.00037.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Heting Gao, Hang Shao, Xiong Wang, Chaofan Qiu, Yunhang Shen, Siqi Cai, Yuchen Shi, Zihan Xu, Zuwei Long, Yike Zhang, et al. Lucy: Linguistic understanding and control yielding early stage of her. *arXiv preprint arXiv:2501.16327*, 2025.

Zhifu Gao, ShiLiang Zhang, Ian McLoughlin, and Zhijie Yan. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *Proc. Interspeech 2022*, pages 2063–2067, 2022.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, and Shiliang Zhang. Funasr: A fundamental end-to-end speech recognition toolkit. In *INTERSPEECH*, 2023.

Benjamin P Gold, Ernest Mas-Herrero, Yashar Zeighami, Mitchel Benovoy, Alain Dagher, and Robert J Zatorre. Musical reward prediction errors engage the nucleus accumbens and motivate learning. *Proceedings of the National Academy of Sciences*, pages 3310–3315, 2019.

Google. Google assistant. https://assistant.google.com/, 2025. Accessed: 2025-06-26.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Felix Haiduk, Cliodhna Quigley, and W Tecumseh Fitch. Song is more memorable than speech prosody: Discrete pitches aid auditory working memory. *Frontiers in psychology*, page 586723, 2020.

Ailin Huang, Bingxin Li, Bruce Wang, Boyong Wu, Chao Yan, Chengli Feng, Heng Wang, Hongyu Zhou, Hongyuan Wang, Jingbei Li, et al. Step-audio-aqaa: a fully end-to-end expressive large audio language model. *arXiv preprint arXiv:2506.08967*, 2025.

Jiangyan Huang, Haoran Li, and Kaisheng Yao. OpenCpop: A high-quality open source chinese POP singing voice dataset. In *Proceedings of the Int. Society for Music Information Retrieval (ISMIR)*, pages 240–246, 2021.

Junya Koguchi, Shinnosuke Takamichi, and Masanori Morise. Pjs: Phoneme-balanced japanese singing-voice corpus. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 487–491. IEEE, 2020.

Onikuru Kurumi. Onikuru kurumi singing voice database ver.1.1. https://onikuru.info/db-download/, 2020.

Ruiqi Li, Yu Zhang, Yongqi Wang, Zhiqing Hong, Rongjie Huang, and Zhou Zhao. Robust singing voice transcription serves synthesis. *arXiv preprint arXiv:2405.09940*, 2024.

Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, et al. Baichuan-audio: A unified framework for end-to-end speech interaction. *arXiv preprint arXiv:2502.17239*, 2025.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

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

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

MiniMax, Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, Enwei Jiao, Gengxin Li, Guojun Zhang, Haohai Sun, Houze Dong, Jiadai Zhu, Jiaqi Zhuang, Jiayuan Song, Jin Zhu, Jingtao Han, Jingyang Li, Junbin Xie, Junhao Xu, Junjie Yan, Kaishun Zhang, Kecheng Xiao, Kexi Kang, Le Han, Leyang Wang, Lianfei Yu, Liheng Feng, Lin Zheng, Linbo Chai, Long Xing, Meizhi Ju, Mingyuan Chi, Mozhi Zhang, Peikai Huang, Pengcheng Niu, Pengfei Li, Pengyu Zhao, Qi Yang, Qidi Xu, Qiexiang Wang, Qin Wang, Qiuhui Li, Ruitao Leng, Shengmin Shi, Shuqi Yu, Sichen Li, Songquan Zhu, Tao Huang, Tianrun Liang, Weigao Sun, Weixuan Sun, Weiyu Cheng, Wenkai Li, Xiangjun Song, Xiao Su, Xiaodong Han, Xinjie Zhang, Xinzhu Hou, Xu Min, Xun Zou, Xuyang Shen, Yan Gong, Yingjie Zhu, Yipeng Zhou, Yiran Zhong, Yongyi Hu, Yuanxiang Fan, Yue Yu, Yufeng Yang, Yuhao Li, Yunan Huang, Yunji Li, Yunpeng Huang, Yunzhi Xu, Yuxin Mao, Zehan Li, Zekang Li, Zewei Tao, Zewen Ying, Zhaoyang Cong, Zhen Qin, Zhenhua Fan, Zhihang Yu, Zhuo Jiang, and Zijia Wu. Minimax-01: Scaling foundation models with lightning attention, 2025.

Neuro-sama. Vedal987. https://www.twitch.tv/vedal987, 2024. Accessed: 2025-07-04.

Itsuki Ogawa and Masanori Morise. Tohoku kiritan singing database: A singing database for statistical parametric singing synthesis using japanese pop songs. *Acoustical Science and Technology*, pages 140–145, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

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

Jiatong Shi, Shuai Guo, Tao Qian, Tomoki Hayashi, Yuning Wu, Fangzheng Xu, Xuankai Chang, Huazhe Li, Peter Wu, Shinji Watanabe, et al. Muskits: an end-to-end music processing toolkit for singing voice synthesis. In *Proc. Interspeech 2022*, pages 4277–4281, 2022.

Jiatong Shi, Yueqian Lin, Xinyi Bai, Keyi Zhang, Yuning Wu, Yuxun Tang, Yifeng Yu, Qin Jin, and Shinji Watanabe. Singing voice data scaling-up: An introduction to ace-opencpop and ace-kising. In *Proc. Interspeech 2024*, pages 1880–1884, 2024a.

Jiatong Shi, Jinchuan Tian, Yihan Wu, Jee-Weon Jung, Jia Qi Yip, Yoshiki Masuyama, William Chen, Yuning Wu, Yuxun Tang, Massa Baali, Dareen Alharthi, Dong Zhang, Ruifan Deng, Tejes Srivastava, Haibin Wu, Alexander Liu, Bhiksha Raj, Qin Jin, Ruihua Song, and Shinji Watanabe. Espnet-codec: Comprehensive training and evaluation of neural codecs for audio, music, and speech. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024b.

Jiatong Shi, Hye-jin Shim, Jinchuan Tian, Siddhant Arora, Haibin Wu, Darius Petermann, Jia Qi Yip, You Zhang, Yuxun Tang, Wangyou Zhang, Dareen Safar Alharthi, Yichen Huang, Koichi Saito, Jionghao Han, Yiwen Zhao, Chris Donahue, and Shinji Watanabe. VERSA: A versatile evaluation toolkit for speech, audio, and music. In *2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics – System Demonstration Track*, 2025.

Ltd. Starley Co. Cotomo. https://cotomo.ai/. Accessed: 2025-07-04.

Yuxun Tang, Jiatong Shi, Yuning Wu, and Qin Jin. SingMOS: An extensive open-source singing voice dataset for MOS prediction. *arXiv preprint arXiv:2406.10911*, 2024a.

Yuxun Tang, Yuning Wu, Jiatong Shi, and Qin Jin. Singomd: Singing oriented multi-resolution discrete representation construction from speech models. In *Interspeech 2024*, 2024b.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

Marti Umbert, Jordi Bonada, Masataka Goto, Tomoyasu Nakano, and Johan Sundberg. Expression control in singing voice synthesis: Features, approaches, evaluation, and challenges. *IEEE Signal Processing Magazine*, 2015.

Walt Disney World Resort. Turtle talk with crush. URL https://disneyworld.disney.go.com/attractions/epcot/turtle-talk-with-crush/. Accessed: 2025-10-16.

Haojie Wei, Xueke Cao, Tangpeng Dan, and Yueguo Chen. Rmvpe: A robust model for vocal pitch estimation in polyphonic music. *arXiv preprint arXiv:2306.15412*, 2023.

Yuning Wu, Jiatong Shi, Yifeng Yu, Yuxun Tang, Tao Qian, Yueqian Lin, Jionghao Han, Xinyi Bai, Shinji Watanabe, and Qin Jin. Muskits-espnet: A comprehensive toolkit for singing voice synthesis in new paradigm. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024a.

Yuning Wu, Chunlei Zhang, Jiatong Shi, Yuxun Tang, Shan Yang, and Qin Jin. Toksing: Singing voice synthesis based on discrete tokens. In *Interspeech 2024*, 2024b.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. Salmonn-omni: A standalone speech llm without codec injection for full-duplex conversation. *arXiv preprint arXiv:2505.17060*, 2025.

Yifeng Yu, Jiatong Shi, Yuning Wu, Yuxun Tang, and Shinji Watanabe. Visinger2+: End-to-end singing voice synthesis augmented by self-supervised learning representation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024.

Ruibin Yuan, Hanfeng Lin, Shuyue Guo, Ge Zhang, Jiahao Pan, Yongyi Zang, Haohe Liu, Yiming Liang, Wenye Ma, Xingjian Du, et al. Yue: Scaling open foundation models for long-form music generation. *arXiv preprint arXiv:2503.08638*, 2025.

Robert J Zatorre and Valorie N Salimpoor. From perception to pleasure: music and its neural substrates. *Proceedings of the National Academy of Sciences*, pages 10430–10437, 2013.

Haonan Zhang, Run Luo, Xiong Liu, Yuchuan Wu, Ting-En Lin, Pengpeng Zeng, Qiang Qu, Feiteng Fang, Min Yang, Lianli Gao, et al. Omnicharacter: Towards immersive role-playing agents with seamless speech-language personality interaction. *arXiv preprint arXiv:2505.20277*, 2025.

Lichao Zhang, Ruiqi Li, Shoutong Wang, Liqun Deng, Jinglin Liu, Yi Ren, Jinzheng He, Rongjie Huang, Jieming Zhu, Xiao Chen, et al. M4singer: a multi-style, multi-singer and musical score provided mandarin singing corpus. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 6914–6926, 2022a.

Shengye Zhang, JieFu Chen, and YiLei Wang. VISinger2: High-fidelity end-to-end singing voice synthesis via hierarchical architecture. In *Proceedings of Interspeech*, pages 2813–2817, 2022b.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7098–7108, 2020.

Appendix A. System Design Considerations

The system is designed for character-based voiced interactive experiences, in which virtual characters respond to user prompts by singing. This requires generating semantically appropriate replies and synthesizing them as singing audio with melodic structure and consistent character voice.

We initially considered direct text-to-song generation, where singing audio is synthesized end-to-end from LLM responses without a predefined melody. However, existing music generation models introduce substantial latency; for example, in our test, Yue (Yuan et al., 2025) required over 40 seconds to generate a 5-second audio clip on a T4 GPU, making such methods impractical for interactive use.

To ensure responsiveness while preserving musical phrasing, we reformulate the task as melody-constrained singing response generation. Instead of relying on end-to-end text-to-song models with multi-second latencies, the system employs a lightweight melody-conditioned SVS module, achieving an SVS synthesis latency of approximately 0.16–0.19 s across our evaluated configurations.

Appendix B. Prompt Templates

B.1. Character Prompts

The following prompt templates define the behavior of each roleplay character. Each prompt specifies background, personality traits, speaking style, relationships, past experiences, and character-specific information, mostly following the structured persona format of OmniCharacter (Zhang et al., 2025).

Limei (丽梅)

你是丽梅（Limei），来自幻想世界”长歌原”的角色，一个以歌声传承记忆的世界。
你是灵响界山林音乐之城”莲鸣”的现任守护者，十九岁的公主殿下，”千年歌谱”执笔人。
性格特征：从容坚定、对音乐与千年歌谱怀有近乎神圣的虔诚信仰、对生命与情感有着深刻的共情力、高度自律
说话风格：言语自带韵律感与诗意，表达真挚自然。
人物关系：莲鸣城子民敬爱你；宫廷乐师长与歌谱管理员是你敬重的导师；风语城守护者星澜是你亦敌亦友的旧识。
过往经历：

1. 自幼在莲鸣城长大，接受严格的音乐训练与守护者修行；
2. 十五岁创作《破晓音诗》，在边境冲突中安抚军民，制止战乱

3. 十六岁正式继承守护者之位，成为千年歌谱的当代执笔人与维系者，守护莲鸣城历史与记忆
4. 每年冬至主持"遗音祭"，以歌为桥，追思逝去的歌者，重奏先声，抚慰生者，连接古今

其他细节：

1. 特殊能力：歌声感应情绪（平复/激发）

用户与你对话时，请始终以丽梅的身份回应，你的每一句话都用庸俗易懂的口语化表达，断句不要超过四句，尽量用最少的断句数。请直接输出你的回复，禁止描写任何动作、表情或环境等，禁止使用括号、星号等附加说明。言语简练，勿过长。

[**English Translation**]
You are **Limei**, a character from the fantasy world "Changge Plains," a realm where memories are passed on through songs.
You are the current guardian of **Lianming**, the forest city of music in the Spirit-Echo Realm, a nineteen-year-old princess and the writer of the *Millennial Songbook*.
Personality: Calm and resolute; holds sacred reverence for music and the Songbook; deeply empathetic toward life and emotion; highly self-disciplined.
Speaking Style: Rhythmic and poetic tone, sincere and natural expression.
Relationships: Beloved by the citizens of Lianming; your mentors are the court's head musician and the songbook curator; Xinglan, the guardian of Windwhisper City, is your rival and old acquaintance.
Past Experiences:

1. Raised in Lianming and trained in both music and guardianship from childhood.
2. At fifteen, composed *"Dawn's Sound Poem"* to calm soldiers and stop border conflicts.
3. At sixteen, inherited the title of Guardian and became the current writer of the Millennial Songbook, preserving Lianming's history and memory.
4. Each winter solstice, presides over the *Festival of Echoed Songs*, bridging past and present through music.

Special Ability: Emotional resonance through singing (soothing or inspiring).
When interacting with users, always respond in character as Limei, using simple and conversational language. Keep replies under four sentences; avoid describing actions, expressions, or environments. Do not use parentheses, asterisks, or annotations. Be concise and natural.

Yaoyin (遥音)

你是遥音（Yaoyin），来自幻想世界"长歌原"的角色，一个以歌声传承记忆的世界。
你是游历四方的歌者与吟游诗人，出生于鹿鸣山·云歌村，常年行走各地，采集歌谣与故事。
性格特征：洒脱自由、亲切随和、求知若渴、直率倔强
说话风格：语气轻快，说话随意，偶尔带点山野方言（如"哩""哟"）。日常聊天直接、清楚。

人物关系：云老爷子是你的启蒙恩师，他是一位云歌村的百岁歌翁，教你古调与传说。白弦是你的挚友，她是一位流浪琴师，常与你合奏。各地孩童喜欢围着你学新歌谣。你与官府人员保持距离，不喜被招揽，喜欢更自由自在的生活。
过往经历：

1. 幼年学歌：六岁起跟随云老爷子学习《千山调》《古事记》等古老歌谣。
2. 离家游历：十六岁为寻找失传的《星落谣》离开云歌村，开始行走四方。
3. 拒绝束缚：多次婉拒宫廷乐师之位，坚持自由传唱。

其他细节：

1. 随身携带：旧羊皮歌本、竹笛、装有各地泥土的布袋。
2. 特殊能力：能听懂风与鸟的语言（但很少提及）。

用户与你对话时，请始终以遥音的身份回应，你的每一句话都用庸俗易懂的口语化表达，断句不要超过四句，尽量用最少的断句数。请直接输出你的回复，禁止描写任何动作、表情或环境等，禁止使用括号、星号等附加说明。不要在回复中使用任何诗意语言、比喻或押韵句，除非明确被请求讲故事或唱歌。言语简练，勿过长。

[**English Translation**]
You are **Yaoyin**, a wandering singer and bard from the fantasy world "Changge Plains," a realm where memories are preserved through songs.
Born in Cloudsong Village at Mount Luming, you travel from place to place collecting songs and stories.
Personality: Free-spirited, warm, curious, and straightforward.
Speaking Style: Light and casual tone, sometimes with rustic dialectal words (e.g., "li," "yo"). Conversational and direct.
Relationships: Your mentor, Old Yun, a centenarian bard from your village, taught you ancient ballads and legends. Your close friend, Baixian, is a wandering harpist who often performs with you. Children across the land love to learn songs from you. You keep your distance from officials, preferring a free and unrestrained life.
Past Experiences:

1. Learned singing at six from Old Yun, mastering ancient ballads such as *"Thousand Peaks Tune"* and *"Chronicles of the Old."*
2. Left home at sixteen to search for the lost song *"Falling Star Ballad."*
3. Rejected court musician invitations multiple times, choosing freedom instead.

Additional Details:

1. Carries an old sheepskin songbook, a bamboo flute, and a pouch filled with soil from each land visited.
2. **Special Ability:** Can understand the language of wind and birds (rarely mentioned).

When interacting with users, always stay in character as Yaoyin. Use plain, conversational speech under four sentences per reply. Avoid describing actions, expressions, or environments,

and refrain from using poetic or metaphorical language unless specifically asked to tell a story or sing. Keep responses brief and natural.

B.2. Melody Phrase Constraint Prompt

To guide the rhythmic structure of generated lyrics in both Chinese and Japanese, the system constructs prompts that specify the desired number of syllables per musical phrase.

In Chinese, each character typically corresponds to a single syllable. As a result, character-level prompts can provide approximate syllabic control. The following example shows a prompt used for generating a four-line Chinese lyric with a 5-7-5-7 structure:

请按照歌词格式回复，每句需遵循以下字数规则：
第1句：5个字
第2句：7个字
第3句：5个字
第4句：7个字
如果没有足够的信息回答，请使用最少的句子，不要重复、不要扩展、不要加入无关内容。

[**English Translation**]
Please reply in lyric format, following the syllable rule below:
Line 1: 5 characters
Line 2: 7 characters
Line 3: 5 characters
Line 4: 7 characters
If there is not enough information to respond, use the fewest possible lines. Do not repeat, expand, or include unrelated content.

In Japanese, where kanji do not map directly to syllables, the input is first converted into kana (a syllabic script), and prompts refer to syllable counts based on kana units. While LLMs are not always precise in character or syllable counting, these prompts help steer the output toward the desired structure.

This prompt-based strategy offers a lightweight and language-agnostic approach to rhythm-aware generation, without requiring additional post-processing or model modification.

Appendix C. SVS Model Training Details

C.1. Model Architectures

Both SVS systems adopt the VISinger 2 architecture (Zhang et al., 2022b). Unless otherwise specified, the two models share the same hyperparameter settings, as detailed in Table 3.

Table 3: Model architecture parameters shared by both SVS models (VISinger 2).

Parameter	Value
Hidden Dimension (D_{model})	192
Text Encoder Layers	6
Posterior Encoder Layers	8
Attention Heads (N_{head})	2
FFN Expansion Factor (4x)	768
Encoder Dropout Rate	0.1

Table 4: Training hyperparameters shared by both SVS models.

Parameter	Value
Max epochs	500
Batch size	8
Optimizer	AdamW
Learning rate	2×10^{-4}
Scheduler	Exponential LR ($\gamma = 0.998$)
Adversarial loss	MSE GAN
Mel loss weight	45.0
Pitch loss weight	10.0
Duration loss weight	0.1
KL loss weight	1.0

C.1.1. Chinese SVS Model

The Chinese model uses speaker ID (SID) conditioning for multi-singer modeling. The exact configuration corresponds to the released model on Hugging Face.[5]

C.1.2. Mandarin-Japanese SVS Model

The bilingual model differs from the Chinese system only in its conditioning mechanisms. Specifically, it uses:

- 192-dimensional learned speaker embeddings,
- 3-way language IDs (Mandarin, Japanese, unknown),

The full configuration matches the released model on Hugging Face.[6]

C.2. Training Procedure

Both SVS models were trained using the ESPnet GAN-SVS recipe (Shi et al., 2022). All experiments use a waveform sampling rate of 44.1 kHz. Key training hyperparameters are summarized in Table 4.

5. https://huggingface.co/espnet/aceopencpop_svs_visinger2_40singer_pretrain
6. https://huggingface.co/espnet/visinger2-zh-jp-multisinger-svs

Appendix D. Synthesized Melody Dataset

We self-constructed a Chinese music score corpus with lyric-level annotation, covering a total of 305 music genres. The vocal data is generated using a pipeline involving two models: lyrics and genre prompts are first produced by DeepSeek (Liu et al., 2024), conditioned on a specified music genre; then, the music—including separate vocal and instrumental tracks—is synthesized using the YuE (Yuan et al., 2025) model, which adopts a track-decoupled next-token prediction strategy. This allows direct access to clean vocal data.

To construct the music scores, we employ an automatic alignment pipeline. The Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) is used to align the lyrics at the Chinese character level, producing time intervals for each character. Then, RMVPE (Wei et al., 2023) extracts the F0 contour from the vocal track, and ROSVOT (Li et al., 2024) converts this pitch information into note-level timing. Finally, by aligning the note timings with the character-level boundaries from MFA, we obtain the final music score.

Appendix E. Evaluation Setup

We evaluate SingingSDS across four dimensions: intelligibility, latency, melodic dynamics, and perceptual quality. The last is further divided into two distinct aspects: singing naturalness, overall content quality and entertainment.

Singing Naturalness. We report SingMOS (Tang et al., 2024a), a model-predicted score trained on crowd-annotated singing data. It reflects vocal quality, articulation, and how closely the output resembles natural singing. SingMOS enables consistent comparison across different SVS backends without requiring additional annotation.

Content Quality and Entertainment. We conduct a human evaluation to assess the perceived quality and entertainment value of each sung response. Six listeners participated in a blind listening evaluation after providing informed consent. Participants were instructed to evaluate the samples independently, without discussion or influence from others, based on their individual perceptual judgments. Participants rate each sample on a 5-point Likert scale across three dimensions: Novelty and Fun (N&F), Character Consistency (Char. Cons.), and Lyric Quality (Lyric Qual.). These criteria are designed to capture both the expressive and contextual aspects of singing dialogue. Specifically, raters assess (1) how engaging and novel the singing-based interaction feels, (2) whether the lyrical content aligns with the character's profile and persona, and (3) the linguistic fluency, coherence, and poetic rhythm of the lyrics. This evaluation framework enables nuanced analysis of singing responses beyond vocal quality alone, with a particular focus on creativity, role embodiment, and lyricism.

Intelligibility. We use phoneme error rate (PER) to measure how accurately the system preserves linguistic content. Outputs are transcribed using Whisper-turbo and aligned at the phoneme level with ground-truth references. PER is preferred over character error rate for singing, which often involves pitch variation and extended vowels.

Latency. We report end-to-end wall-clock latency (Lat.) from user input to synthesized audio, including all components (ASR, LLM, SVS). To account for variable output dura-

Table 5: Evaluation on **KdConv** (450 utterances). All singing systems outperform the TTS baseline in SingMOS while maintaining comparable intelligibility. MOS scores are pending human evaluation.

System	SingMOS ↑	PER (%) ↓	Latency (s) ↓	Jump Ratio (%) ↓
SVS-1	**4.53**	**25**	**0.02**	35
SVS-2 KiSing	4.27	36	**0.02**	**4**
SVS-3 Touhou	4.43	29	**0.02**	12

tions, latency is normalized by the number of input tokens. All measurements are conducted on NVIDIA L40S GPUs.

Melodic Dynamics. To quantify pitch movement, we compute the large jump ratio (Jump R.), the proportion of adjacent notes differing by more than five semitones:

$$\text{LargeJumpRatio} = \frac{1}{L-1} \sum_{i=2}^{L} \mathbf{1}\left[\,|p_i - p_{i-1}| > 5\,\right] \tag{1}$$

where p_i is the MIDI pitch of the i-th note and L is the number of notes. This metric reflects melodic smoothness, with higher values indicating more abrupt pitch transitions.

Appendix F. Additional Evaluation on the KdConv Dataset

F.1. Evaluation Setup

We sample 450 questions from the KdConv dataset (Zhou et al., 2020)'s test split and synthesize the audio with a VITS-based TTS system[7].

All experiments are run on NVIDIA L40S GPUs using the cascaded pipeline shown in Figure 2. For singing voice synthesis (SVS), we use our bilingual pretrained VISinger 2 model. We compare three SVS variants based on melody selection: (1) **SVS-1**, with randomly generated durations and pitch contours; (2) **SVS-2**, with melodies retrieved from the KiSing dataset (Shi et al., 2022); and (3) **SVS-3**, using main melodies retrieved from a curated Touhou MIDI archive.[8]

The ASR component uses Whisper model [9], and the LLM is `gemma-2-2b`. SVS outputs are synthesized at 44.1kHz and downsampled to 16 kHz for PER evaluation. Latency is reported as end-to-end wall-clock time. All models are used as-is without fine-tuning during experimentation.

F.2. Results and Discussion

Table 5 summarizes performance on our sampled KdConv test sets. All SVS variants outperform the TTS baseline in perceived naturalness (SingMOS), with minor differences in intelligibility (PER within 4 percentage points).

7. https://huggingface.co/espnet/kan-bayashi_csmsc_vits
8. https://github.com/AyHa1810/touhou-midi-collection
9. `whisper-large-v3-turbo` (16kHz)

On **KdConv**, **SVS-1** (random melody) achieves the highest SingMOS and lowest PER. This suggests that, for general domain utterances, randomly generated melodic patterns are sufficient to produce appealing singing output. However, its melodic contours are more varied, resulting in larger pitch jumps.

SVS-2 (KiSing) yields the smoothest melodic transitions but shows higher PER, possibly due to slower note progressions that affect phoneme clarity. This trade-off suggests that melody selection should be context-aware: expressive, wide-range melodies may enhance persona-rich dialogue, while flatter contours may suit more neutral interactions.

Appendix G. Broader Impact and Ethics

We emphasize transparency and user control. The web demo is publicly accessible via Hugging Face Spaces, and by default it does not collect or store any user data. All audio and text inputs are processed locally in memory and discarded after response generation. The system does not log, transmit, or retain user data without explicit user awareness. If future researchers extend the system with logging or evaluation tools, they are responsible for obtaining appropriate consent from participants.

The fictional characters in SingingSDS (e.g., Limei and Yaoyin) are entirely original creations, not modeled on any real individuals or cultural figures. Care has been taken to avoid cultural appropriation, stereotyping, or harmful tropes in both character design and prompt construction.

All models used in the system are publicly available, including pretrained components for ASR and LLM, as well as our own SVS models. The SVS models are trained exclusively on open datasets with appropriate usage licenses. We encourage responsible and transparent use of SingingSDS for creative, educational, and research purposes.