Make-A-Voice: Revisiting Voice Large Language Models as Scalable Multilingual and Multitask Learners

Rongjie Huang[1*], **Chunlei Zhang**[2*], **Yongqi Wang**[1*], **Dongchao Yang**[3], **Jinchuan Tian**[4], **Zhenhui Ye**[1], **Luping Liu**[1], **Zehan Wang**[1], **Ziyue Jiang**[1], **Xuankai Chang**[4], **Jiatong Shi**[4], **Chao Weng**[2], **Zhou Zhao**[1†], **Dong Yu**[2]

Zhejiang University[1], Tencent AI Lab[2],

The Chinese University of Hong Kong[3], Carnegie Mellon University[4]

{rongjiehuang,cyanbox,zhaozhou}@zju.edu.cn

{cleizhang,dyu}@global.tencent.com

Abstract

Large language models (LLMs) have successfully served as a general-purpose interface across multiple tasks and languages, while the adaptation of voice LLMs is mostly designed for specific purposes (either single-task or monolingual), where the advantages of LLMs especially for low-resource language processing and zero-shot task generalization are less exploited in the audio community. To bridge the gap, we introduce Make-A-Voice as a multimodal voice LLM and conduct a comprehensive study on its capability to deal with multiple tasks/languages. When trained on ∼200K hours of 6-language data for 4 voice generation applications, Make-A-Voice emerges notable advantages: 1) as scalable learners to improve performance with end-to-end local and global multiscale transformers; and 2) as multitask learners by adjusting prompts to share common knowledge across modalities (speech/singing) and present in-context learning abilities by generalizing to unseen tasks not explicitly train on; 3) as multilingual learners to alleviate data scarcity of low-resource languages by including rich-resource language training data. Experimental results demonstrate that Make-A-Voice exhibits superior audio quality and style similarity compared with competitive baseline models in monolingual/cross-lingual voice generation. [1]

1 Introduction

Large language models (LLMs) (Devlin et al., 2018; Raffel et al., 2020; Ouyang et al., 2022; Zhang et al., 2022b) trained on massive corpora of texts have shown their ability to perform new tasks from textual instructions or a few examples. The LLM-based interfaces excel at generating text for tasks that require the modeling of complex interactions, and they are trained to predict sequences of discrete tokens, have also been adapted to continuous, audio signals (Borsos et al., 2022; Agostinelli et al., 2023; Kreuk et al., 2022).

Current voice LLMs (Kharitonov et al., 2023; Wang et al., 2023a; Zhang et al., 2023) cast voice synthesis as a language modeling task in a discrete representation space. VALL-E (Wang et al., 2023a) proposes a language model approach for TTS with audio codec codes as intermediate representations. Kharitonov et al. (2023) introduce a hierarchical approach that combines two types of audio tokens. Zhang et al. (2023) train a multilingual conditional codec language model to predict the acoustic token sequences of speech in different languages. Despite the success achieved, most existing voice LLMs are designed for specific purposes (single-task or monolingual), where the advantages of LLMs especially for low-resource language processing and zero-shot task generalization are less exploited in the audio community.

To bridge the gap, we leverage the intuition that GPTs can process multiple tasks or languages as a general-purpose interface and emerge strong in-context learning capabilities. In this work, we conduct a comprehensive study on voice LLMs' capability to deal with multiple tasks (i.e., speech and singing voice modeling) and multiple languages (i.e., rich and low resource data). We refer to this resulting model as Make-A-Voice, for **m**ultimodal LLMs to synthesize and manipulate **voice** signals. Make-A-Voice is a decoder-only model trained on a mixture of tasks, emerging capabilities of cross-task knowledge sharing, generalization to new tasks, and alleviating data scarcity of low-resource languages.

Make-A-Voice employs self-supervised tokens for the unified voice generation pipeline: 1) Semantic tokens determine the semantic meaning given text or speech; 2) Acoustic tokens provide

[*] Equal contributions

[†] Corresponding author

[1] Audio samples are available at https://M-Voice.github.io

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10929–10942

acoustic information given various control conditions, which can be learned in a large amount of self-supervised audio-only data. Make-A-Voice is trained on ∼200K hours of multilingual data in 6 languages, and we introduce 4 applications: text-to-speech (TTS), voice conversion (VC), singing voice synthesis (SVS), and singing voice conversion (SVC). Experimental results demonstrate that Make-A-Voice achieves SOTA results in monolingual/cross-lingual zero-shot voice generation. Subjective and objective evaluation show that Make-A-Voice exhibits superior audio quality and style similarity compared with baselines. The key takeaways are as follows:

- **Make-A-Voice as scalable learners to improve performance.** Make-A-Voice inherently leverages the transformer architecture and scales the models' size in depth and width. Make-A-Voice predicts long sequences with end-to-end differentiable multiscale (local and global) transformers to reduce the extremely long sequences of the acoustic token, where scaling the model size (520M (medium), and 1.2B (large) parameter) from 160M (base) results in 15% and 37% improvement for TTS.

- **Make-A-Voice as multilingual learners to alleviate data scarcity of low-resource languages.** Combined with more rich-resource language data, model training benefits from a variety of acoustic conditions on top of semantic meanings. Make-A-Voice excels at zero-shot transferring acoustic attributes (speaker identity, emotion, prosody) even in low-resource languages, which has witnessed a 20% SIM gain and 7.9% WER drop in TTS synthesis.

- **Make-A-Voice as multitask learners for cross-task knowledge sharing, even generalizable to unseen tasks.** Make-A-Voice is trained with a combination of semantic and acoustic modeling tasks across speech/singing voice modalities by adjusting prompts (referring to Figure 1); 1) It shares common knowledge across tasks (e.g., multi-quantization codec modeling across modalities), as evidenced by 27% SIM improvement in SVS with the help of speech tasks; and 2) Make-A-Voice illustrates in-context learning abilities by generalizing to tasks not explicitly trained on, including cross-lingual timbre transferring, generating coherent emotion, and noise continuations.

2 Related Works

2.1 Generative Voice Models

Text-guided voice synthesis (text-to-speech and singing voice synthesis) typically converts input text into mel-spectrogram (e.g., Tacotron (Wang et al., 2017), FastSpeech (Ren et al., 2019)), which is then transformed to waveform using a separately trained vocoder (Kong et al., 2020; Huang et al., 2021). Recent generative models cast voice synthesis as a language modeling task to perform in-context learning: VALL-E (Wang et al., 2023a) use discrete codes derived from an off-the-shelf neural audio codec model, and regard TTS as a conditional language model. Zhang et al. (2023) leverage back-translation and prompt-guided LLMs for high-quality TTS with limited supervision. Jiang et al. (2023) train a prosody language model with arbitrary-length speech prompts to produce expressive and controlled prosody.

Despite the success achieved, most voice LLMs are designed for single-task or as monolingual. In this work, Make-A-Voice presents emerging capabilities as **scalable multilingual and multitask learners**, which provides critical takeaways for the.

2.2 Multitask Learning

Building a simple and multitask learning framework has attracted increasing attention in the community: NANSY (Choi et al., 2021a, 2022) is trained in a self-supervised manner that does not require any annotations paired with audio, and efficiently tackles multiple applications after training the backbone network. Lee et al. (2021) design a multitask learning framework with joint speech and text training that enables the model to generate dual mode output (speech and text) simultaneously in the same inference pass. Wang et al. (2023b) combine neural codec language modeling with multitask learning using task-dependent prompting, capable of zero-shot TTS and various speech transformation tasks, dealing with clean and noisy signals. Rubenstein et al. (2023) fuse text-based and speech-based language models into a unified multitask architecture to process and generate text and speech. In this work, we demonstrate voice LLMs' capabilities as **multitask learners** across various voice generation tasks. Make-A-Voice demonstrates the improved capability with multitask learning by **transferring common knowledge across modalities** and illustrates in-context learning abilities by performing tasks Make-A-Voice is **not explicitly**

trained on.

2.3 Multilingual Learning

Multilinguality has been a very active research area in speech and NLP. Duquenne et al. (2022) introduce a large-scale multilingual speech-to-speech corpus and demonstrate that model pre-training and sparse scaling using a mixture of experts bring significant gains to translation performance. VoiceBox (Le et al., 2023) trains a non-autoregressive flow-matching model on 50K hours of multilingual audiobooks from six languages. Massively Multilingual Speech (Pratap et al., 2023) increases supported languages with rich-resource datasets and effective self-supervised learning. In this work, we present voice LLMs' capabilities as **multilingual learners**, where the rich-resource language data is included to learn various acoustic conditions, alleviating the data scarcity of prompt-guided in-context learning for low-resource language.

3 Voice Large Language Models

In this section, we overview the discrete voice representation, namely semantic and acoustic tokens, and then introduce the decoder-only unified voice synthesis model Make-A-Voice. Next, we introduce the designs of the multitask learning in Section 3.3 and multilingual approaches in Section 3.4, as well as the scalable multi-scale transformer architecture in Section 3.5.

3.1 Voice Representation

Semantic tokens. It is crucial to extract rich linguistic information from the speech signal. To this end, we resort to XLSR-53: a wav2vec 2.0 model pre-trained on 56k hours of speech in 53 languages (Conneau et al., 2020). In the following, a k-means algorithm is applied to the learned representations of the unlabelled speech to generate K_1 cluster centroids at every 20-ms frame. In the end, a speech utterance y is represented as semantic tokens with $[s_1, s_2, \ldots, s_T], s_i \in \{0, 1, \ldots, K_1 - 1\}, \forall 1 \leq i \leq T$, where T is the number of frames.

Acoustic tokens. The audio encoder E of codec models (Zeghidour et al., 2021; Défossez et al., 2022) consists of several convolutional blocks with a total downsampling rate of 320 and generates continuous representations at every 20-ms frame in 16kHz. The residual vector-quantizer Q produces discrete representations a_q with a codebook size of K_2, using a vector quantization layer (Vasuki and Vanathi, 2006). In the end, we **flatten all the codebooks** and thus a speech utterance y is represented as acoustic tokens with $[a_1, a_2, \ldots, a_T], a_i \in \{0, 1, \ldots, K_2 - 1\}, \forall 1 \leq i \leq T$, where T is the number of frames.

3.2 Make-A-Voice: Controllable Voice LLMs

As illustrated in Figure 1 and Table 1, Make-A-Voice casts voice synthesis as language modeling tasks with self-supervised tokens, where voice synthesis is broken down into more manageable pieces (i.e., semantic modeling or acoustic modeling) and jointly learned in a decoder-only language model, where various conditioning mechanisms are investigated in 1) **semantic modeling**: semantic tokens s determine the semantic meaning given text or speech; 2) **conditional acoustic modeling**: acoustic tokens a are guided by control conditions (speaker, emotion, prosody, and style) and learned on top of semantic meanings in a large amount of self-supervised audio-only data. In the end, a unit-based vocoder synthesizes high-fidelity waveforms from compressed acoustic representations.

- **Zero-shot TTS / VC.** Given a target text $\mathbf{y}$, TTS models first determine semantic tokens s, and then perform in-context learning given acoustic prompt $\mathbf{a_p}$ derived from a reference utterance. During training, we randomly select two non-overlapping speech windows from each example and consider one window as a prompt and the other as a target. For VC, we extract the semantic tokens from the Hubert with the K-means model.

- **Zero-shot SVS / SVC.** Different from speech, singing voice requires accurate rhythm and pitch control guided by MIDI representation, where the fundamental frequency $\mathbf{F_0}$ and phone-level duration are given respectively in semantic and acoustic modeling. In practice, $\mathbf{F_0}$ could be predicted by a separately-trained neural network provided MIDI score, and thus we directly take the $\mathbf{F_0}$ value as condition signals for simplification following (Liu et al., 2022).

3.3 Multitask Learner

As illustrated in Table 1, Make-A-Voice presents strong controllability with flexible conditioning as a unified voice synthesis framework, where Make-A-Voice is trained with a combination of semantic and acoustic modeling tasks across speech/singing voice modalities by adjusting prompts. We signal to the model which task it should perform on a given input by prefixing the information with a

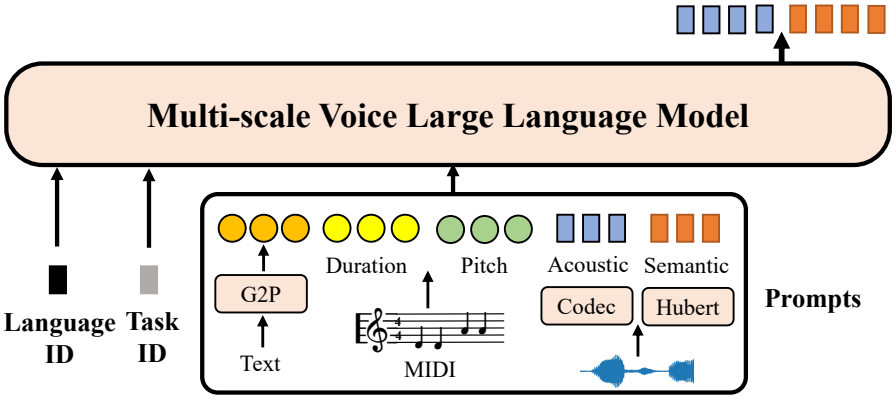


Figure 1: Voice generation tasks are jointly learned in a decoder-only language model. Prompts can be adjusted for different tasks with a variety of conditions (speaker, emotion, prosody, and style).

Task	Modality	Prompts
Semantic modeling		
1	Speech	Text
2	Singing	Text, duration
Acoustic modeling		
3	Speech	Reference acoustic, semantic, pitch $\mathbf{F}_0$
4	Singing	Reference acoustic, semantic

Table 1: Multitask learning by adjusting prompts.

tag specifying the task, referring to 3.4. Make-A-Voice, as a multitask learner, exhibits the following competitive advantages:

- Towards a general-purpose interface across various voice generation tasks.
- Cross-task knowledge sharing: multi-task learning improves overall performance by transferring and sharing common knowledge (e.g., multi-quantization codec modeling).
- Generalization to new tasks: it presents in-context learning abilities by performing tasks Make-A-Voice is **not explicitly trained on**, such as cross-lingual timbre transferring, generating coherent emotion, and noise continuations.

As expected, Make-A-Voice demonstrates the outperformed audio quality and style similarity in zero-shot speech and singing voice synthesis, where a complex task is broken down into more manageable pieces. We refer the reader to Section 5.3 for a detailed analysis of our findings.

3.4 Multilingual Learner

With large-scale training data and powerful models, speech models can now generate high-quality samples with unseen styles (e.g., timbre, emotion, and prosody) derived from an acoustic reference (i.e., custom voice). However, replicating this success is a significant challenge for low-resource languages due to data scarcity.

To alleviate it, Make-A-Voice includes rich-resource language and trains a model on a mixture of arbitrarily multilingual voice, where 1) limited **low-resource language** data connects its text to semantic meanings, and 2) a large amount of **rich-resource language** data contains many speakers with various accents, diverse demographics, and heterogeneous recording conditions, which introduces a variety of acoustic conditions on top of semantic meanings with different conditioning mechanisms. As such, voice LLMs demonstrate capturing acoustic diversity (speaker identity, emotion, prosody) in zero-shot scenarios, even for low-resource languages. We refer the reader to Section 5.3 for a summary of our findings.

We signal the model which language to perform on a given input by prefixing the input with a tag specifying the task and language. For example, to query the model to perform text-to-semantic translation on an utterance in English, the tokenized input would be preceded by the two tags [En] [T2S]. To enable the model to be cross-lingual, we employ **multilingual HuBERT and codec models** to respectively extract the semantic and acoustic discrete representations, which are pre-trained on human voice in multiple languages.

3.5 Scalable Architecture

Recent research (Agostinelli et al., 2023; Kreuk et al., 2022) leverages the transformer architecture (Vaswani et al., 2017) for improving scalability and proposes to represent audio signals as multiple streams n_q of discrete tokens and flatten these codes to the length of $T \times n_q$, where T is the number of frames. It comes at a high computational cost for extremely long sequences due to the quadratic cost of self-attention and large feed-forward networks per position.

To alleviate it, Make-A-Voice (denoted as θ_{AR}) predicts long sequences with end-to-end differen-

Tasks	Language	Dataset	Testing set
Group by tasks (Make-A-Voice as multitask learners.)			
Speech generation/conversion	Ja, De, Fr, En, Es, Zh	Librilight, Gigaspeech, WenetSpeech, CSS...	LibriTTS/VCTK
Singing generation/conversion	En, Zh	OpenSinger, M4Singer, CSD...	Opencpop
Group by languages (Make-A-Voice as multilingual learners.)			
Low-resource	Ja, De, Fr, Es	CSS, CSD	CSS
Rich-resource	En, Zh	OpenSinger, Librilight, Gigaspeech...	Opencpop, LibriTTS

Table 2: Dataset usage in training and inference stages. We have attached detailed information on the data configuration in Appendix A.

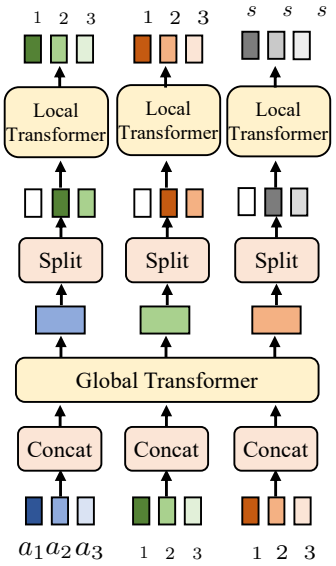


Figure 2: Overview of the architecture of differentiable multi-scale transformer.

tiable multiscale transformers similar to Yu et al. (2023); Yang et al. (2023). This enables sub-quadratic self-attention, unlocking better performance at reduced cost for both training and generation. As illustrated in Figure 2: 1) the token embedding matrix E_G maps integer-valued tokens $x_{0..T}$ to m dimensional embeddings, and concatenate with continuous speech representation in time axis (if any), following which 2) we chunk it into patches of size P of length $K = \frac{T}{P}$, 3) a large global transformer $\theta_{AR}^{\text{global}}$ module outputs patch representations $\mathbf{G_o^{1:K}} = \theta_{AR}^{\text{global}}(\mathbf{G_i^{0:K-1}})$, and 4) a small local transformer module operates on a single patch containing P elements, each of which is the sum of an output from the global model and an embedding of the previous tokens, and autoregressively predict the next patch $\mathbf{L_o^{1:K}} = \theta_{AR}^{\text{local}}(\mathbf{L_i^{0:K-1}} + \mathbf{G_o^{1:K}})$.

Make-A-Voice presents the improvements primarily from scaling the models' size in depth and width without the requirement of scattered model-specific methodologies. As expected, scaling the model size (160M (base), 520M (medium), and 1.2B (large) parameter) results in better scores. We refer the reader to Section 5.3 for our findings.

3.6 Reconstructing High-Fidelity Waveforms

We train a unit-based neural vocoder from scratch for the acoustic unit to waveform generation. Inspired by BigVGAN (Lee et al., 2022), the synthesizer includes the generator and multi-resolution discriminator (MRD). The generator is built from a set of look-up tables (LUT) that embed the discrete representation and a series of blocks composed of transposed convolution and a residual block with dilated layers. The transposed convolutions upsample the encoded representation to match the input sample rate. Details are included in Appendix C.2.

4 Training and Evaluation

4.1 Dataset

Table 2 lists the used datasets with six languages, with **English (En), Chinese (Zh)** (totally ∼200k hours) as rich-resource language settings and four languages **French (Fr), German (De), Spanish (Es), Japanese (Ja)** (totally ∼80 hours) as low-resource settings. Overall, we have ∼**200k** hours of 16 kHz audio as training data. For text sequence, we tokenize it into the phoneme sequence with an open-source grapheme-to-phoneme conversion tool (Sun et al., 2019). During the evaluation, we randomly choose sentences to construct the zero-shot testing set for each application task, in which the voice used for prompting is never seen by the model at training, and it has to reproduce the characteristics from a single prompt example. We have attached detailed data configuration in Appendix A.

4.2 Evaluation Metrics

Intelligibility and accuracy. We employ word error rate (WER) to evaluate the intelligibility of the generated speech by transcribing it using a wav2vec ASR system. We transcribe the translated speech for accuracy and then calculate the BLEU score (Papineni et al., 2002) between the generated and the reference text. For English-only setups, we use the large model[2] pretrained and fine-tuned

[2] https://github.com/facebookresearch/fairseq/blob/main/examples/wav2vec/README.md#

on Libri-Light and Librispeech on 16kHz sampled speech audio. For multilingual settings, we use ASR models publicly released on HuggingFace following (Duquenne et al., 2022).

Style quality and similarity. Speaker similarity score (SIM) assesses the coherence of the generated speech in relation to the speaker's characteristics, which is calculated as the cosine similarity between the speaker embeddings of the generated speech and the desired speech signals. F0 Frame Error (FFE) measures the timbre and prosody similarity of synthesized and reference audio, respectively.

Subjective evaluation. We also conduct a crowd-sourced human evaluation via Amazon Mechanical Turk, which is reported with 95% confidence intervals (CI), and analyze two aspects: style similarity (speaker, emotion, and prosody) and audio quality (clarity, high-frequency), respectively scoring SMOS and MOS. More information has been attached in Appendix D.

4.3 Baseline

We compare the generated audio samples with other systems, including 1) GT, the ground-truth audio; 2) YourTTS (Casanova et al., 2022), GenerSpeech (Huang et al., 2022b), VALL-E (Wang et al., 2023a) for English zero-shot TTS; 3) YourTTS (Casanova et al., 2022) for zero-shot multilingual TTS; 4) NANSY (Choi et al., 2022) and PPG-VC (Liu et al., 2021) for VC; 5) Diffsinger (Liu et al., 2022) and FFT-Singer for SVS;

4.4 Model Configurations

For semantic representations, we apply XLSR-53 pre-trained on 56k hours of speech in 53 languages (Conneau et al., 2020) and use k-means to discretize 12th-layer embeddings into semantic tokens with a codebook of size 1000 and a total downsampling rate of 320. For acoustic representation, we train the SoundStream model with 12 quantization levels, each with a codebook of size 1024 and the same downsampling rate of 320. We take three quantization levels as the acoustic tokens, representing each frame as a flat sequence of tokens from the first, second, and third quantization layers. We trained three sets of Make-A-Voice, with 160M (base), 520M (medium), and 1.2B (large) parameters. As for the unit-based vocoder, we use the modified V1 version of BigVGAN. A comprehensive table of hyperparameters is available in Appendix B. Except explicitly stated, we use our 520M (medium) model for downstream evaluation.

During training, we train Make-A-Voice for 100K steps using 8 NVIDIA V100 GPUs with a batch size of 6000 tokens for each GPU on the publicly-available *fairseq* framework (Ott et al., 2019). Adam optimizer is used with $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-9}$. S_3 model is optimized with a segment size of 8192 and a learning rate of 1×10^{-4} until 500K steps using 4 NVIDIA V100 GPUs. For sampling, we employ top-p (Holtzman et al., 2019) sampling with p = 0.25.

5 Results And Analysis

5.1 Make-A-Voice as Multitask Learners

Zero-shot Text-to-Speech. 1) For the intelligibility of the generated speech, Make-A-Voice has achieved a WER of 6.7, comparable with other systems, indicating that Make-A-Voice could generate accessible speech of good quality as previous non-autoregressive TTS families. 2) For audio quality, Make-A-Voice has achieved the highest MOS with scores of 4.04 compared with the baseline models, demonstrating the effectiveness of the vocoder in generating high-fidelity waveforms. 3) Regarding style similarity, Make-A-Voice scores the SIM of 0.85, showing that Make-A-Voice surpasses the state-of-the-art models in transferring the style of custom voices. Informally, Make-A-Voice is optimized in a large amount of self-supervised data, which contains many speakers with various accents, diverse demographics, and heterogeneous recording conditions, to improve robustness and generalization in zero-shot scenarios.

Using the examples provided on its demo page, we also compare Make-A-Voice with VALL-E in a small-scale subjective test. We synthesize utterances using the same transcripts and prompts and conduct the objective and subjective test with the same protocol described above. Table 3 shows that, in these examples, Make-A-Voice obtains 1.5 lower in WER and 0.04 higher in SIM than baseline models in zero-shot synthesis.

Singing Voice Synthesis. Table 5 demonstrates that Make-A-Voice (SVS) outperforms the baseline system by a large margin in terms of pitch similarity, showing distinct 70%/30% superiority over FFT-Singer/DiffSinger in terms of FFE objective evaluation. Make-A-Voice can resemble the note prompt and demonstrates its precise pitch reconstruction. Regarding singer similarity, Make-

pre-trained-models

Model	MOS (↑)	SMOS (↑)	WER (↓)	SIM (↑)
GT	4.23±0.09	/	4.1	/
GenerSpeech	3.99±0.08	3.77±0.08	8.6	0.83
YourTTS	3.89±0.08	3.72±0.06	12.1	0.78
Make-A-Voice	**4.04±0.07**	**3.81±0.08**	6.7	**0.85**
Small-Scale Subjective Test				
VALL-E	3.92±0.12	3.81±0.07	4.5	0.79
Make-A-Voice	4.01±0.06	3.87±0.04	3.0	0.83

Table 3: Quality and style similarity of generated samples in zero-shot text-to-speech.

Model	TTS-WER	TTS-SIM	VC-SIM
Base	9.8	0.84	0.75
Medium	8.3	0.86	**0.76**
Large	**6.1**	**0.87**	**0.76**

Table 4: LJSpeech results for different model sizes, namely 160M (base), 520M (medium), and 1.2B (large) parameter models. We investigate voice large language models as scalable learners.

Model	MOS (↑)	SMOS (↑)	SIM (↑)	FFE (↓)
GT	4.08±0.08	/	/	/
FFT-Singer	3.86±0.05	3.91±0.08	0.66	0.12
Diffsinger	3.96±0.07	3.94±0.07	0.67	0.11
Make-A-Voice (Zero-shot)	**3.99±0.06**	**3.96±0.05**	**0.78**	**0.08**

Table 5: SVS. Note that FFT-Singer and Diffsinger conduct in-domain generation with seen speaker while Make-A-Voice presents zero-shot SVS.

Model	MOS (↑)	SMOS (↑)	SIM (↑)
Voice Conversion			
Prompt	4.26±0.06	/	/
NANSY	3.89±0.08	3.73±0.10	0.68
PPG-VC	3.97±0.06	**3.82±0.05**	0.78
Make-A-Voice (Zero-shot)	**4.02±0.08**	3.78±0.06	**0.80**
Singing Voice Conversion			
Prompt	4.21±0.05	/	/
Make-A-Voice	3.96±0.06	3.72±0.05	**0.76**

Table 6: Zero-shot VC and SVC.

A-Voice scores the highest SIM of 0.78, surpassing the state-of-the-art models in transferring the style of custom singing voices in zero-shot scenarios even though the voice used for prompting is never seen at training.

Voice Conversion and Singing Voice Conversion. Table 6 shows that Make-A-Voice scores the comparable overall SIM of 0.93 with baseline. It excels at converting speaker identity even in a zero-shot scenario, attributing to the scalable training data covering diverse speakers with various accents. For audio quality, it presents high perceptual quality with outperformed MOS evaluation. To conclude, Make-A-Voice converts the timbre with better naturalness and comparable speaker similarity to baseline models, even though the model is trained without any text transcript paired with audio recordings. For singing voice conversion (SVC), Make-A-Voice also excels at converting singer identity and presents good perceptual quality and naturalness.

5.2 Make-A-Voice as Multilingual Learners

Table 7 presents cross-lingual zero-shot TTS results, where the audio context and the target text are in different languages. For each target text, we sample one 3-second-long audio context from each language, which creates language transfer directions in total. Compared with YourTTS, Make-A-Voice yields better results in most languages, obtaining lower WER and higher SIM averaged across audio contexts. Regarding low-resource language, Make-A-Voice presents potential improvement for the limited usage of training data at this time.

5.3 Analysis and Ablation Studies

To verify the emerging capabilities of Make-A-Voice as scalable multilingual and multitask learners, we conduct ablation studies and discuss the key findings as follows. In this section, we first analyze the model scalability, then investigate the benefits of multilingual and multitask training, and finally explore its generalization to unseen tasks.

Scalability to improve performance. Table 4 reports LJSpeech results for different model sizes, namely 160M (base), 520M (medium), and 1.2B (large) parameter models. As expected, scaling the model size results in better scores. However, this comes at the expense of longer training and inference time. Increasing the model size from 520M to 1.2B leads to additional gains of a further 40% reduction in WER for TTS tasks with a similar style similarity.

Multilingual Learning to alleviate data scarcity of low-resource languages. Data scarcity is a significant challenge to replicating the success of voice LLM for low-resource languages. To verify the effectiveness of multilingualism in alleviating data scarcity, we train Make-A-Voice using 16hrs **De** data and another one with a combination of six-language data; Make-A-Voice leverages a joint vocabulary, where rich-resource language data contains many speakers with various accents, diverse demographics, and heterogeneous recording conditions, introducing a variety of acoustic conditions on top of semantic meanings, especially

	Prompt	De		En		Es		Fr	Zh	Ja
		WER	SIM	WER	SIM	WER	SIM	SIM	SIM	SIM
YT	De	6.0	0.81	3.1	0.71	4.1	0.71	0.72	0.74	/
	En	8.0	0.79	6.3	0.71	3.0	0.78	0.71	0.72	/
	Es	10.3	0.71	2.6	0.73	12.3	0.80	0.70	0.70	/
	Fr	12.7	0.76	5.1	0.71	18.6	0.67	0.79	0.67	/
	Zh	21.3	0.72	11.0	0.65	2.0	0.75	0.76	0.75	/
	Ja	6.1	0.80	10.1	0.73	2.1	0.69	0.82	0.79	/
	AVG	**10.7**	**0.76**	6.3	0.70	**7.0**	0.73	0.75	0.72	/
Ours	De	10.1	0.78	4.2	0.75	14.1	0.75	0.78	0.72	0.70
	En	15.1	0.78	9.1	0.77	9.1	0.80	0.74	0.79	0.67
	Es	13.0	0.76	7.1	0.75	13.0	0.78	0.78	0.68	0.70
	Fr	22.0	0.70	3.6	0.71	11.0	0.73	0.78	0.77	0.69
	Zh	10.3	0.68	5.3	0.71	18.1	0.69	0.76	0.70	0.68
	Ja	9.1	0.79	8.0	0.77	8.1	0.85	0.80	0.76	0.93
	AVG	13.2	0.75	**6.2**	**0.74**	12.2	**0.77**	**0.77**	**0.74**	**0.72**

Table 7: Quality and style similarity of generated samples in multilingual zero-shot text-to-speech. YT refers to YourTTS. We report SIM for simplification for the Fr, Zh, and Ja languages.

in low-resource languages. Table 8 shows the improved performance with the combination of arbitrarily multilingual voice, leading to the 7.9 reduction in WER and 0.2 SIM gain for text-to-speech synthesis.

Task	MOS (↑)	WER (↓)	SIM (↑)
Multilingual Learning			
TTS	3.82±0.06	18.0	0.65
TTS (M)	**3.93±0.05**	**10.1**	**0.78**
VC	3.83±0.06	/	0.52
VC (M)	**3.91±0.05**	/	**0.72**
Multitask Learning			
SVS	3.90±0.08	/	0.61
SVS (M)	**3.99±0.06**	/	**0.78**

Table 8: We investigate voice LLMs as multitask and multilingual learners. For multilingual settings, Make-A-Voice (TTS/VC) is trained in **De** combined with rich-resource language data. For multitask settings, Make-A-Voice (SVS) is trained jointly with speech tasks. **M:** multilingual or multitask learning.

Multitask learning for cross-task knowledge sharing. As illustrated in Table 8, we observe that joint training with speech tasks has witnessed the gains of 0.09 improvement in MOS and a 0.17 point improvement in SIM for singing voice synthesis. Since common knowledge can be shared across tasks (such as multi-layer quantization codec modeling), SVS distinctly benefits from large-scale speech tasks.

Multitask learning for generalizing to tasks not explicitly trained on. Besides quantitative results, we present in-context learning abilities by performing tasks **not explicitly trained on**, including cross-lingual timbre transferring, generating coherent emotion, and noise continuations. We have attached the information on testing data in Appendix A. As shown in the demo page, we find that 1) Make-A-Voice can preserve the **emotion** in the prompt at a zero-shot setting, even if the model is not fine-tuned on an emotional TTS dataset; 2) Make-A-Voice effectively reproduces the characteristics from a **cross-lingual** style prompt, which has not been seen during training; and 3) In a noisy environment, the model also presents the acoustic consistency and maintain the **noise conditions** from the prompt.

6 Conclusion

To bridge the gap where multiple tasks and languages are less exploited in voice LLMs compared to GPTs, we introduced Make-A-Voice, a multimodal LLM to synthesize and manipulate voice signals. Make-A-Voice took self-supervised tokens first to determine the semantic meaning and then learned acoustic information given condition signals, and we conducted a comprehensive study on its capabilities to deal with multiple tasks and languages. When trained on ∼200K hours of 6-language data for 4 voice generation applications, Make-A-Voice emerged notable advantages as 1) **as scalable learners to improve performance** with end-to-end differentiable local and global multiscale transformers; and 2) **as multitask learners** to share common knowledge across modalities (speech/singing) and presented in-context learning abilities by performing tasks not explicitly trained

on; 3) **as multilingual learners to alleviate data scarcity of low-resource languages** by including rich-resource language training data. Experimental results demonstrated that Make-A-Voice achieved state-of-the-art results in monolingual/cross-lingual zero-shot voice generation. The subjective and objective evaluation showed that Make-A-Voice exhibited superior audio quality and similarity compared with baselines.

7 Limitation and Potential Risks

Although Make-A-Voice as a voice LLM is successfully applied to multilingual zero-shot voice signals at scale, it still suffers from some limitations: 1) Make-A-Voice introduces a strong dependency on the quality of the audio tokenizer. 2) The model only shows in-context learning ability on voice synthesis, rather than all voice recognition and understanding tasks, and 3) a longer sequence length typically requires more computational resources, and degradation could be witnessed with decreased training data.

The low-resource scenario refers to the lack of labeled training data, and we construct a total of around 80 hours of labeled data for low-resource downstream tasks. We leave the study of truly resource-limited language LLMs for future works.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62222211.

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*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2022. Audiolm: a language modeling approach to audio generation. *arXiv preprint arXiv:2209.03143*.

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.

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

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

Hyeong-Seok Choi, Jinhyeok Yang, Juheon Lee, and Hyeongju Kim. 2022. Nansy++: Unified voice synthesis with neural analysis and synthesis. *arXiv preprint arXiv:2211.09407*.

Soonbeom Choi, Wonil Kim, Saebyul Park, Sangeon Yong, and Juhan Nam. 2021b. CSD: Children's Song Dataset for Singing Voice Research. Important: version 1.1 is available.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

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

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*.

Paul-Ambroise Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswani, Changhan Wang, Juan Pino, Benoît Sagot, and Holger Schwenk. 2022. Speechmatrix: A large-scale mined corpus of multilingual speech-to-speech translations. *arXiv preprint arXiv:2211.04508*.

Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, et al. 2021. Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario. *arXiv preprint arXiv:2104.03603*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Rongjie Huang, Feiyang Chen, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. 2021. Multi-singer:

Fast multi-singer singing voice vocoder with a large-scale corpus. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3945–3954.

Rongjie Huang, Chenye Cui, Feiyang Chen, Yi Ren, Jinglin Liu, Zhou Zhao, Baoxing Huai, and Zhefeng Wang. 2022a. 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, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. 2022b. Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech synthesis. *arXiv preprint arXiv:2205.07211*.

Ziyue Jiang, Yi Ren, Zhenhui Ye, Jinglin Liu, Chen Zhang, Qian Yang, Shengpeng Ji, Rongjie Huang, Chunfeng Wang, Xiang Yin, et al. 2023. Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias. *arXiv preprint arXiv:2306.03509*.

Jacob Kahn, Morgane Riviere, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. IEEE.

Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. 2023. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *arXiv preprint arXiv:2302.03540*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Proc. of NeurIPS*.

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

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. *arXiv preprint arXiv:2306.15687*.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, et al. 2021. Direct speech-to-speech translation with discrete units. *arXiv preprint arXiv:2107.05604*.

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*.

Songxiang Liu, Yuewen Cao, Disong Wang, Xixin Wu, Xunying Liu, and Helen Meng. 2021. Any-to-many voice conversion with location-relative sequence-to-sequence modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1717–1728.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*, 32.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Hao Sun, Xu Tan, Jun-Wei Gan, Hongzhi Liu, Sheng Zhao, Tao Qin, and Tie-Yan Liu. 2019. Token-level ensemble distillation for grapheme-to-phoneme conversion. *arXiv preprint arXiv:1904.03446*.

A Vasuki and PT Vanathi. 2006. A review of vector quantization techniques. *IEEE Potentials*, 25(4):39–47.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. 2017. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 6:15.

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

Xiaofei Wang, Manthan Thakker, Zhuo Chen, Naoyuki Kanda, Sefik Emre Eskimez, Sanyuan Chen, Min Tang, Shujie Liu, Jinyu Li, and Takuya Yoshioka. 2023b. Speechx: Neural codec language model as a versatile speech transformer. *arXiv preprint arXiv:2308.06873*.

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*.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

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*.

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*.

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.

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

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.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022b. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*.

A Data

In this section, we describe details of the data usage in training and evaluating Make-A-Voice.

- Common Voice (Ardila et al., 2019) consists of text paired with recordings where people were asked to read the text aloud.
- Librilight (Kahn et al., 2020) contains 60K hours of unlabelled speech from audiobooks in English, and LibriSpeech (Panayotov et al., 2015), LibriTTS (Zen et al., 2019), Gigaspeech (Chen et al., 2021), AISHELL (Fu et al., 2021), VCTK (Veaux et al., 2017) datasets include transcriptions.
- CSD (Choi et al., 2021b) contains multilingual singing voice. We also use the female-singer OpenCPOP (Wang et al., 2022), multi-singer dataset OpenSinger (Huang et al., 2021), and M4Singer (Zhang et al., 2022a) as the singing voice data.

B Model Configurations

We list the model hyper-parameters of Make-A-Voice in Table 9.

C Applications

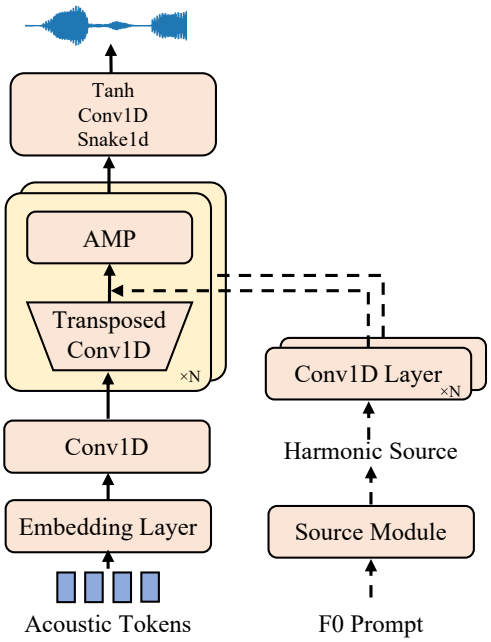


Figure 3: Overview of the unit-based vocoder. The F0 auxiliary input denoted with dotted lines is included only in singing voice synthesis.

C.1 MIDI-to-F0 Converter

Singing voice synthesis (SVS) is a task that generates singing voices from the given music score and lyrics like human singers. Following (Liu et al., 2022; Zhang et al., 2022a), the SVS system typically includes the MIDI-to-F0 converter to predict F0 explicitly. Though the SVS system can be further improved with the direct MIDI condition and implicit F0 prediction, this is beyond our focus.

C.2 Unit-based Vocoder

The generator of the unit-based vocoder is built from a set of look-up tables (LUT) that embed the discrete representation, and a series of blocks composed of transposed convolution and a residual block with dilated layers. We train the enhanced vocoder with the weighted sum of the least-square adversarial loss, the feature matching loss, and the spectral regression loss on mel-spectrogram, where the training objective formulation and hyperparameters follow Kong et al. (2020); Lee et al. (2022).

For speech generation, we train the vocoder with only the discrete unit sequences as input. For singing voice generation, we further include F0-driven source excitation to stabilize long-continuous waveforms generation following (Liu et al., 2022; Huang et al., 2022a).

As illustrated in Table 10, replacing the unit-based vocoder with a SoundStream decoder for voice synthesis has witnessed a distinct degradation of perceptual quality, proving that the codebook mismatch for the SoundStream decoder between training (12 quantization levels) and inference (3 levels) hurts reconstruction performance. In contrast, a neural vocoder could refine the coarse-grained acoustic tokens and generate waveforms with increasing details.

D Evaluation

D.1 Subjective Evaluation

For audio quality evaluation, 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 style similarity evaluation, we explicitly instruct the raters to "*(focus on the similarity of the*

Hyperparameter		Make-A-Voice
Make-A-Voice Global Base	Transformer Layer	16
	Transformer Embed Dim	768
	Transformer Attention Headers	12
	Number of Parameters	114 M
Make-A-Voice Global Medium	Transformer Layer	20
	Transformer Embed Dim	1152
	Transformer Attention Headers	16
	Number of Parameters	320 M
Make-A-Voice Global Large	Transformer Layer	24
	Transformer Embed Dim	1536
	Transformer Attention Headers	32
	Number of Parameters	930 M
Make-A-Voice Local	Transformer Layer	6
	Transformer Embed Dim	Same as global
	Transformer Attention Headers	8
	Number of Parameters	46/101/303 M
BigVGAN Vocoder	Upsample Rates	[5, 4, 2, 2, 2, 2]
	Hop Size	320
	Upsample Kernel Sizes	[9, 8, 4, 4, 4, 4]
	Number of Parameters	121.6M

Table 9: Hyperparameters of Make-A-Voice.

Table 10: Ablation studies.

Model	STOI (↑)	MCD (↓)
S_3: SoundStream	0.92	1.90
S_3: Unit Vocoder	**0.93**	**1.56**

style (timbre, emotion, and prosody) to the reference, and ignore the differences of content, grammar, or audio quality.)". In the SMOS (similarity mean opinion score) tests, we paired each synthesized utterance with a ground truth utterance to evaluate how well the synthesized speech matches that of the target speaker. Each pair is rated by one rater.

Our subjective evaluation tests are crowdsourced and conducted by 20 native speakers via Amazon Mechanical Turk. The screenshots of instructions for testers have been shown in Figure 5. We paid $8 to participants hourly and totally spent about $600 on participant compensation. A small subset of speech samples used in the test is available at https://M-Voice.github.io/.

D.2 Objective Evaluation

Cosine similarity is an objective metric that measures speaker similarity among multi-speaker audio. We compute the average cosine similarity between embeddings extracted from the synthesized and ground truth embeddings to measure the speaker similarity performance objectively.

Word Error Rate (WER) evaluates the faithfulness to the input transcript by transcribing the synthesized utterances using a wav2vec ASR system.

F0 Frame Error (FFE) combines voicing decision error and F0 error metrics to capture F0 information.

Mel-cepstral distortion (MCD) measures the spectral distance between the synthesized and reference mel-spectrum features.

Short-time objective intelligibility (STOI) assesses the denoising quality for speech enhancement.

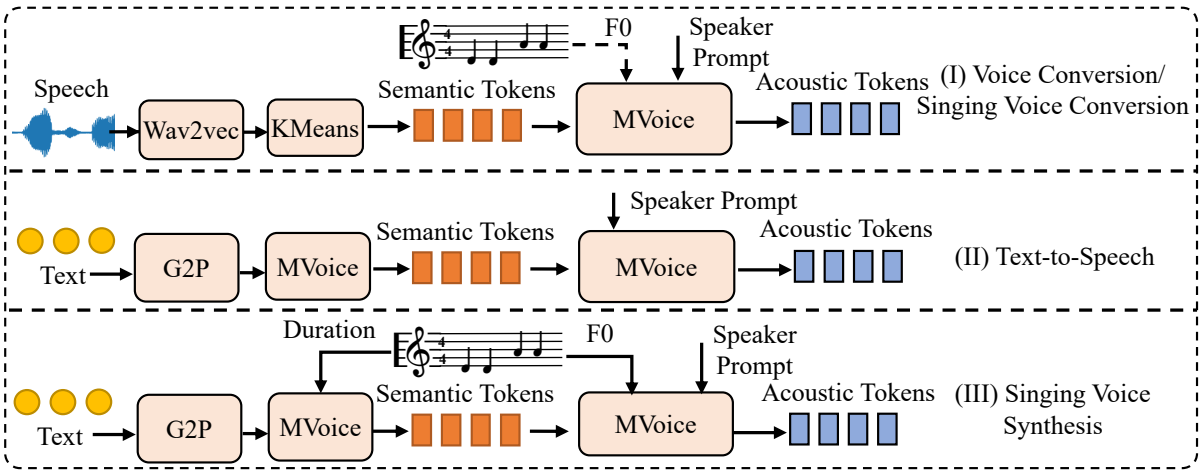
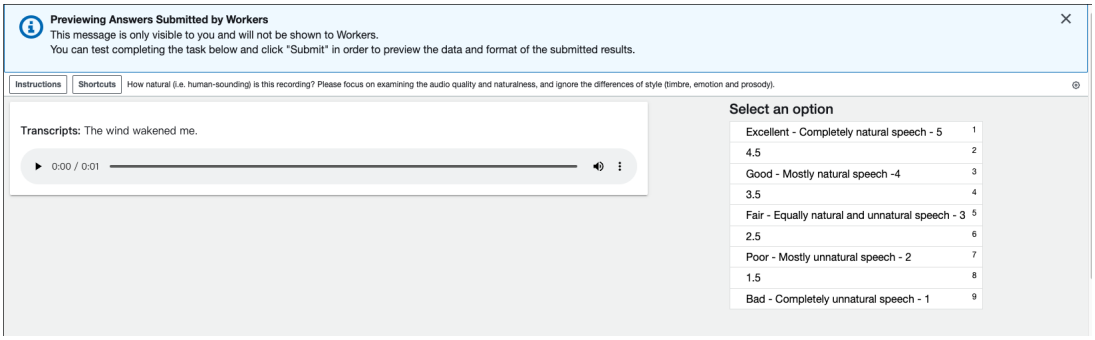
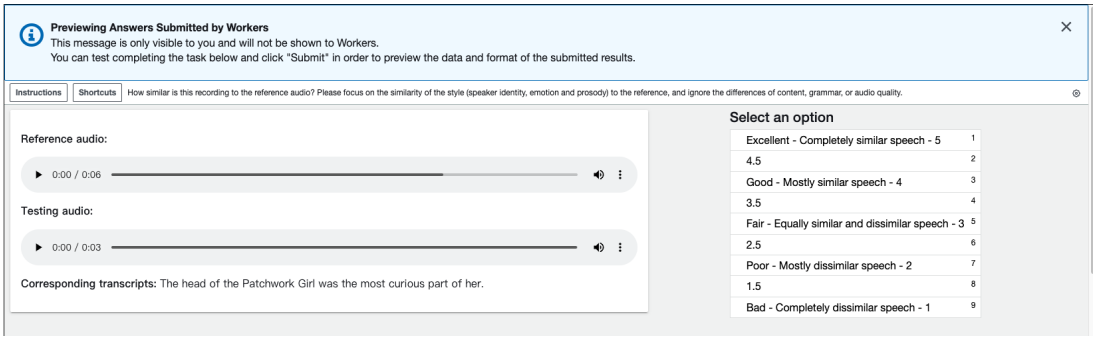


Figure 4: We introduce 4 exemplar applications, including voice conversion (VC), text-to-speech (TTS), singing voice synthesis (SVS), singing voice conversion (SVC), that can be tackled by sharing a voice synthesis framework with semantic and acoustic tokens.



(a) Screenshot of MOS testing.



(b) Screenshot of SMOS testing.

Figure 5: Screenshots of subjective evaluations.