Speak Foreign Languages with Your Own Voice: Cross-Lingual Neural Codec Language Modeling

Ziqiang Zhang* **Long Zhou*** **Chengyi Wang** **Sanyuan Chen** **Yu Wu** **Shujie Liu**
Zhuo Chen **Yanqing Liu** **Huaming Wang** **Jinyu Li** **Lei He** **Sheng Zhao** **Furu Wei**
Microsoft
https://github.com/microsoft/unilm

Abstract

We propose a *cross-lingual neural codec language model*, VALL-E X, for cross-lingual speech synthesis. Specifically, we extend VALL-E [Wang et al., 2023] and train a multi-lingual conditional codec language model to predict the acoustic token sequences of the target language speech by using both the source language speech and the target language text as prompts. VALL-E X inherits strong in-context learning capabilities and can be applied for zero-shot cross-lingual text-to-speech synthesis and zero-shot speech-to-speech translation tasks. Experimental results show that it can generate high-quality speech in the target language via just one speech utterance in the source language as a prompt while preserving the unseen speaker's voice, emotion, and acoustic environment. Moreover, VALL-E X effectively alleviates the foreign accent problems, which can be controlled by a language ID. Audio samples are available at https://aka.ms/vallex.

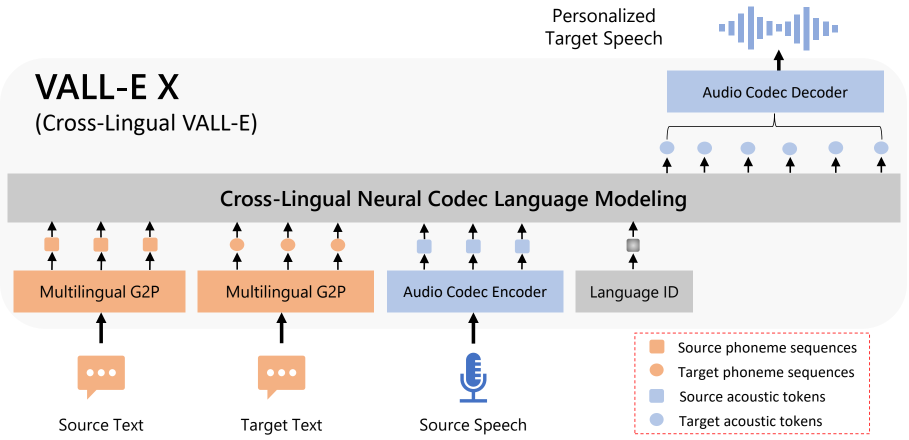

Figure 1: The overall framework of VALL-E X, which can synthesize personalized speech in another language for a monolingual speaker. Taking the phoneme sequences derived from the source and target text, and the source acoustic tokens derived from an audio codec model as prompts, VALL-E X is able to produce the acoustic tokens in the target language, which can be then decompressed to the target speech waveform. Thanks to its powerful in-context learning capabilities, VALL-E X does not require cross-lingual speech data of the same speakers for training, and can perform various zero-shot cross-lingual speech generation tasks, such as cross-lingual text-to-speech synthesis and speech-to-speech translation.

[*] Both authors contributed equally to this work. Correspondence: {lozhou,shujliu,fuwei}@microsoft.com

1 Introduction

Recent years have witnessed significant advancements in end-to-end text-to-speech (TTS) synthesis, and the quality of synthesized speech is even close to human parity [Li et al., 2019, Ren et al., 2019, Tan et al., 2022]. However, these models can only generate high-quality speech for a specific speaker in a specific language. Cross-lingual speech synthesis is a new emerging task that aims to transfer the speaker's voice from one language to another. The speech quality for cross-lingual speech synthesis, especially the speaker similarity, is far behind the monolingual TTS models due to two reasons, 1) data scarcity, as it is difficult to collect multi-lingual speech data for the same speaker, and 2) model capacity, as conventional cross-lingual TTS models are not powerful enough to transfer the speaker voice, speech background, and speaker emotion from the source language speech to the target language speech.

Previous methods to tackle these challenges typically augment end-to-end TTS models with specific subnets for speaker and language control [Nachmani and Wolf, 2019, Zhang et al., 2020, Yang and He, 2020, Ellinas et al., 2022, Cai et al., 2023]. For example, based on the multi-speaker TTS model, Nachmani and Wolf [2019] introduce multiple encoders for each language and additional loss to keep the speaker's identity. Zhang et al. [2020] employs a phonemic representation to capture cross-language information and an adversarial network to disentangle speaker identities. Yang and He [2020] incorporate speaker and language networks with speaker and language IDs as input to deal with the multi-speaker and cross-lingual problems respectively. Yang and He [2022] further propose a multi-task learning method with additional tasks of speaker similarity and language identification. Moreover, Cai et al. [2023] investigates cross-lingual multi-speaker text-to-speech synthesis with sufficient or limited bilingual speech training data. However, the above methods fail to effectively extend to zero-shot scenarios for synthesizing target speech from the unseen source speaker, and often suffer from low speaker similarity and L2 (second-language, or foreign) accent problems [Zhang et al., 2019, Lee et al., 2022].

Table 1: A comparison between VALL-E X and previous cross-lingual TTS systems.

	Previous Systems	**VALL-E X**
Intermediate representation	Mel spectrogram	Audio codec codes
Training data	$<$ 13K hours	70K hours
Speech accent	Foreign	Native
Speaker similarity	Relative low	High
In-context learning	✗	✓
Zero-shot cross-lingual TTS	✗	✓

In this work, we present a novel approach to address these issues by proposing a simple yet effective cross-lingual neural codec language model, VALL-E X, which leverages strong in-context learning capacities to achieve high-quality zero-shot cross-lingual speech synthesis. Based on the knowledge learned from large-scale multi-lingual speech data, VALL-E X is able to transfer the speech characteristics, including the speaker's voice, emotions, and the speech background, from the source language to the target language, and also alleviate the foreign accent problems. More specifically, we first obtain the multi-lingual speech-transcription data from existing ASR data or pseudo-labeled speech data. Then we convert the transcriptions to phoneme sequences with a rule-based converter (G2P tool) and the speech data to acoustic tokens with an offline neural codec encoder. Finally, we concatenate the paired phoneme and acoustic token sequences of each language and train a multi-lingual conditional language model. As illustrated in Figure 1, after training, VALL-E X can predict the acoustic tokens of the target language prompted by the phoneme sequences of both languages and the acoustic tokens of the source language. The generated acoustic token sequence is decompressed to the target speech waveform by an offline audio codec decoder. VALL-E X is trained on two large-scale multi-speaker datasets[1], LibriLight [Kahn et al., 2020] and WenetSpeech [Zhang et al., 2022a], containing about 60,000 hours of English audiobook speech data and 10,000+ hours of multi-domain Chinese ASR data, respectively. The combination of LibriLight and WenetSpeech makes a large multi-lingual multi-speaker multi-domain unclean speech dataset, which significantly

[1] To our knowledge, the largest publicly available speech datasets for English and Chinese.

improves the coverage of different speakers and enhances VALL-E X's generalization capacity. The comparison between VALL-E X and the previous cross-lingual TTS systems are listed in Table 1.

We conduct experiments on two kinds of cross-lingual speech generation tasks, zero-shot cross-lingual text-to-speech synthesis (XTTS), and zero-shot speech-to-speech translation (S2ST). For cross-lingual text-to-speech synthesis, the proposed VALL-E X is evaluated with LibriSpeech [Panayotov et al., 2015] and EMIME [Wester, 2010] for English and Chinese respectively, including English TTS prompted by Chinese speakers and Chinese TTS prompted by English speakers. For zero-shot speech-to-speech translation, EMIME [Wester, 2010] dataset is used for the evaluation of VALL-E X on bidirectional Chinese↔English translation tasks, and it contains bilingual audio recordings by the same speakers. We evaluate the proposed VALL-E X framework from several aspects, including speaker similarity, speech quality (ASR- WER or BLEU), speech naturalness, and human evaluation (e.g., SMOS, MOS, and CMOS). Specifically, due to the strong in-context learning capability, VALL-E X achieves a higher speaker similarity score than the previous SOTA model for the unseen speaker. By training on large-scale speech-transcription data, the proposed VALL-E X significantly reduces the word error rate from 8.53 to 4.07 in the cross-lingual English TTS task, obtains the substantial gain of 3.17 BLEU scores than the strong baseline in S2ST tasks, and achieves better speech naturalness. Furthermore, the human evaluation shows that our VALL-E X outperforms strong baselines in terms of SMOS (4.00 vs. 3.42 in XTTS, 4.12 vs. 3.06 in S2ST), CMOS (+0.24 for ours vs. baseline in XTTS), and MOS (3.87 vs. 3.81 in S2ST). Our contributions can be summarized as follows:

- We develop a cross-lingual neural codec language model VALL-E X with large multi-lingual multi-speaker multi-domain unclean speech data. VALL-E X is a conditional cross-lingual language model predicting the target language acoustic tokens with the source language speech and target language text as prompts.
- The multi-lingual in-context learning framework enables VALL-E X to generate cross-lingual speech maintaining the unseen speaker's voice, emotion, and speech background, prompted by only one sentence in the source language.
- Based on the learned cross-lingual speech modeling ability with the introduced language ID, VALL-E X can generate speech in a native tongue for any speaker and can significantly reduce the foreign accent problem, which is a well-known problem in cross-lingual speech synthesis tasks.
- We apply VALL-E X to zero-shot cross-lingual text-to-speech synthesis and zero-shot speech-to-speech translation tasks. Experiments show that the proposed VALL-E X can beat the strong baseline in terms of speaker similarity, speech quality, translation quality, speech naturalness, and human evaluation.

We encourage readers to listen to the audio samples on our demo page: `https://aka.ms/vallex`.

2 Related Work

Speech/Audio Synthesis With the rapid development and application of neural networks, speech and audio synthesis have made tremendous progress with different network frameworks, such as WaveNet [Oord et al., 2016], HiFi-GAN [Kong et al., 2020a], and Diffwave Kong et al. [2020b]. Academic and industrial communities also pay increasing attention to synthesizing speech or sound from text, namely text-to-speech (TTS) [Li et al., 2019, Ren et al., 2019] or text-to-sound [Yang et al., 2022, Kreuk et al., 2022]. Recently, it is emerging to apply discrete audio representation learning to audio synthesis, e.g., AudioGen [Kreuk et al., 2022] and AudioLM [Borsos et al., 2022]. AudioGen, consisting of an audio encoder, a text encoder, a Transformer encoder, and an audio decoder, is an autoregressive audio generation model with textual descriptions as inputs. AudioLM reviews high-quality audio generation as unidirectional language modeling. In AudioLM, the input audio is mapped to semantic tokens using w2v-BERT [Chung et al., 2021] and acoustic tokens using SoundStream [Zeghidour et al., 2021]. Through three subsequent stages, AudioLM can accomplish speech continuation, acoustic generation, unconditional generation tasks, and so on. The most related work to ours is VALL-E [Wang et al., 2023], which was recently proposed to utilize a neural codec language model to achieve monolingual text-to-speech synthesis. Trained on large-scale speech data, VALL-E shows a strong in-content learning capability and can synthesize high-quality personalized speech prompted by a short recording of an unseen speaker. Different from the above work, this paper focuses on cross-lingual speech synthesis, and the goal is to retain the source language speaker's voice in the synthesized speech of the target language.

Cross-Lingual TTS In cross-lingual speech synthesis, the goal is to synthesize the speech of another language for a monolingual speaker, which is more challenging than conventional monolingual TTS [Nachmani and Wolf, 2019, Zhang et al., 2020, Yang and He, 2022, Cai et al., 2023]. By using shared phonemic input representation across languages and incorporating an adversarial objective to disentangle the speaker's identity and speech content, Zhang et al. [2019] is able to achieve cross-lingual voice cloning within limited speakers. Liu and Mak [2020] also investigate the cross-lingual speech synthesis with speakers' voices enrolled in their native language. In this system, they achieve it using a Tacotron-based synthesizer with a speaker encoder module and introduce a shared phoneme set with IPA to enhance the cross-lingual capability. Aiming at improving the speaker similarity between the synthesized speech and the recordings of the native speaker, authors in Yang and He [2022] propose multi-task learning by jointly training speaker classification and cross-lingual TTS models. Cai et al. [2023] explores cross-lingual multi-speaker speech synthesis under the scenarios of sufficient and limited bilingual training data. In the data-limited scenario, they employ a series of modules including a linguistic feature classifier, a speaker representation extractor, a non-autoregressive multi-speaker voice conversion module, and a neural vocoder, to achieve cross-lingual synthesis. Although previous work has made considerable achievements in cross-lingual TTS, they still suffer from the issue of low speaker similarity and the lack of zero-shot ability. In contrast, leveraging large-scale multi-lingual multi-speaker ASR data, our proposed framework with a neural codec language model demonstrates a strong in-context learning ability to alleviate the above issues.

Speech to Speech Translation (S2ST) S2ST aims to translate the speech of one language to the speech of another language. The initial research and application mainly focus on cascaded S2ST systems [Lavie et al., 1997, Nakamura et al., 2006, Wahlster, 2013], consisting of speech recognition (ASR), machine translation (MT), and speech synthesis (TTS) models. Recently, end-to-end S2ST models have been explored [Jia et al., 2019, Lee et al., 2021a, Jia et al., 2021, Lee et al., 2021b, Wei et al., 2022, Huang et al., 2022, Li et al., 2022], achieving the direct conversion from source speech to target speech. However, there is still an unsolved problem to reserve the source sound characteristics (e.g. speaker, emotion, and the speech background) in generated speech. This challenge is largely due to the zero-shot nature as the bilingual speech data from the same speakers are hard to collect. Though researchers have put much effort into constructing speech-to-speech translation corpora, such as Voxvopule [Wang et al., 2021], CVSS [Jia et al., 2022b], and SpeechMatrix [Duquenne et al., 2022], they are either synthesized from text or mined from multilingual speech corpora thus can not meet the requirement that bilingual data come from the same speakers. At the same time, Translatotron [Jia et al., 2019] tries to synthesize target speech conditioned by the speaker embedding extracted from the source speech, but it misses richer voice information due to the limitation of the speaker embedding. Translatotron 2 [Jia et al., 2021] retrains the speaker voices relying on the pseudo bilingual speech data of the same speakers generated by multi-speaker TTS systems, while the synthetic speech does not completely simulate the speech of the real world. To address these challenges, we propose to equip the cross-lingual neural codec language model with translation modules and show its zero-shot capability to reserve the sound characteristics in the S2ST task.

3 Cross-Lingual Codec Language Model

In this section, we will first present the background, namely conditional codec language model VALL-E, and then introduce the framework of VALL-E X, followed by the multi-lingual training and cross-lingual inference approaches.

3.1 Background

Our VALL-E X is the cross-lingual version of text-to-speech synthesizer VALL-E [Wang et al., 2023], which was recently proposed to leverage a neural codec language model to achieve text-to-speech synthesis. Unlike conventional TTS methods that adopt the continuous regression task, e.g., mel-spectrogram generation, VALL-E regards TTS as a conditional language modeling task with neural codec codes, i.e. acoustic tokens, as an intermediate representation of speech. VALL-E

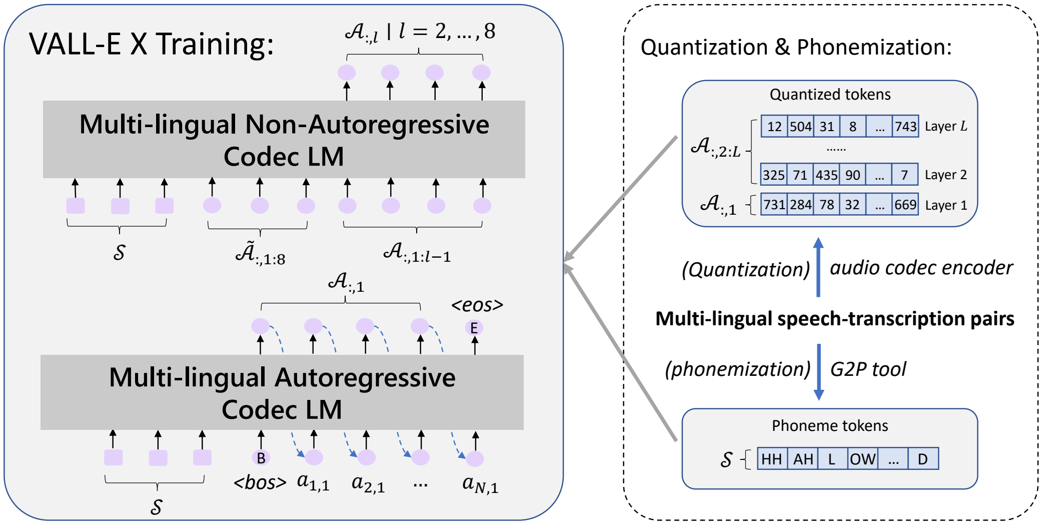
VALL-E X Training:
$\mathcal{A}_{:,l} \mid l = 2, \dots, 8$
Multi-lingual Non-Autoregressive Codec LM
$\mathcal{S}$
$\tilde{\mathcal{A}}_{:,1:8}$
$\mathcal{A}_{:,1:l-1}$
$\mathcal{A}_{:,1}$
<eos>
E
Multi-lingual Autoregressive Codec LM
B
<bos>
$a_{1,1}$
$a_{2,1}$
...
$a_{N,1}$
$\mathcal{S}$
Quantization & Phonemization:
Quantized tokens
$\mathcal{A}_{:,2:L}$
12 504 31 8 ... 743 Layer L
......
325 71 435 90 ... 7 Layer 2
$\mathcal{A}_{:,1}$
731 284 78 32 ... 669 Layer 1
(Quantization)
audio codec encoder
Multi-lingual speech-transcription pairs
(phonemization)
G2P tool
Phoneme tokens
$\mathcal{S}$
HH AH L OW ... D

Figure 2: Training illustration of the cross-lingual neural codec language model VALL-E X, consisting of a multi-lingual autoregressive codec LM (ϕ_{MAR}) and a multi-lingual non-autoregressive codec LM (ϕ_{MNAR}). Multi-lingual acoustic tokens ($\mathcal{A}$) and phoneme sequences ($\mathcal{S}$) are converted from speech and transcription using an audio codec encoder and G2P tool, respectively. During training, we use paired $\mathcal{S}$ and $\mathcal{A}$ from different languages to optimize these two models.

employs two-stage modeling, which first generates the codec codes of the first quantizer of EnCodec [Défossez et al., 2022] from the paired phoneme sequences using an autoregressive language model, and then generates the codes of the rest quantizers in parallel using a non-autoregressive model. After training on the large-scale English speech-transcription dataset LibriLight, VALL-E shows strong in-context learning capabilities. It can generate personalized speech by taking only a 3-second speech fragment as a prompt. Based on VALL-E, our VALL-E X extend to train a cross-lingual neural codec language model, enabling zero-shot cross-lingual capability and supporting cross-lingual TTS or speech-to-speech translation tasks.

3.2 Model Framework

Inspired by VALL-E, the cross-lingual codec language model VALL-E X (denoted as ϕ) leverages a multi-lingual autoregressive codec LM and a multi-lingual non-autoregressive codec LM to generate acoustic tokens at different granularities, as shown in the left part of Figure 2. We also adopt the neural codec model EnCodec [Défossez et al., 2022] as the acoustic quantizer, which is an encoder-decoder model with L quantization layers. We choose $L = 8$ in our experiments, for each layer it produces quantized codes of 1024 entries at 75Hz.

Multi-lingual Autoregressive Codec LM The multi-lingual autoregressive codec LM ϕ_{MAR} is a unidirectional Transformer decoder that autoregressively generates acoustic tokens based on the semantic tokens (phoneme sequence). To make the sentence-level training efficient and accelerate the decoding during inference, similar to VALL-E, the cross-lingual autoregressive codec LM ϕ_{MAR} is only used to predict the acoustic tokens from the first quantizer of EnCodec model.

Formally, based on paired speech-transcription data in any language, let $\mathcal{S}$ denote the transcribed phoneme sequence, and $\mathcal{A}_{:,1} \triangleq \{a_{i,1} | i = 1, \dots, N\}$ denotes the first-layer acoustic tokens extracted from the speech $\mathcal{X}$. The decoder ϕ_{MAR}, modeling the concatenated sequence $\langle \mathcal{S}, \mathcal{A}_{:,1} \rangle$, is trained to predict $\mathcal{A}_{:,1}$ autoregressively. It is optimized by maximizing the log-likelihood,

$$\mathcal{L}_{\text{MAR}} = -\log p_{\text{AR}} \left(\mathcal{A}_{:,1} \mid \mathcal{S}; \phi_{\text{MAR}} \right) = -\log \prod_{i=1}^{N} p \left(a_{i,1} \mid \langle \mathcal{S}, \mathcal{A}_{<i,1} \rangle ; \phi_{\text{MAR}} \right) \tag{1}$$

where $\langle\rangle$ means sequence concatenation operation, and $p(.)$ is the softmax function.

Multi-lingual Non-Autoregressive Codec LM Instead of the autoregressive generation pattern, multi-lingual non-autoregressive codec LM ϕ_{MNAR} is a non-autoregressive Transformer language model aiming at iteratively generating the rest layers of acoustic tokens from the first layer. It is prompted by the phoneme sequence of the current sentence ($\mathcal{S}$) and the acoustic token sequence of another sentence with the same speaker ($\tilde{\mathcal{A}}$). Here $\tilde{\mathcal{A}}$ is taken from the previous sentence in the dataset where the adjusted sentences are usually segmented from the same paragraph. It is expected to have the same characteristics of voice (speaker, speed, background, etc) as the current sentence and is used as an additional reference for cloning the target voice. Like VALL-E, for generating acoustic tokens of each layer $l \in [2, 8]$, the embeddings of $l-1$ layers' acoustic tokens ($\mathcal{A}_{:,1:l-1}$) are summed up layerwise as input. The learning objective for the l-layer acoustic tokens $\mathcal{A}_{:,l}$ can be calculated as

$$\mathcal{L}_{\mathrm{MNAR}} = \sum_{l=2}^{8} \log p_{\mathrm{NAR}} \left(\mathcal{A}_{:,l} \mid \left\langle \mathcal{S}, \tilde{\mathcal{A}}_{:,1:8}, \mathcal{A}_{:,1:l-1} \right\rangle ; \phi_{\mathrm{MNAR}} \right) \tag{2}$$

where $\langle\rangle$ means the sequence concatenation. $p_{\mathrm{NAR}}(.)$ computes the pointwise probabilities of $\mathcal{A}_{:,l}$.

3.3 Multi-lingual Training

In order to learn cross-lingual acoustic conversion information for cross-lingual TTS and speech-to-speech translation tasks, we take advantage of bilingual speech-transcription (ASR) corpus[2], pairs of ($\mathcal{S}^s$, $\mathcal{A}^s$) and ($\mathcal{S}^t$, $\mathcal{A}^t$) to train our multi-lingual codec LMs ϕ_{MAR} and ϕ_{MNAR}, where s and t represent two different (source and target) languages.

Language ID Module Following multi-lingual TTS, we leverage a language ID to guide the speech generation for specific languages in VALL-E X. On the one hand, without language ID, VALL-E X may be confused to select suitable acoustic tokens for the specific language since it is trained with multi-lingual data. On the other hand, some languages have very different characteristics, for example, Chinese is a tone language while English is a non-tone language, which increases the difficulty of adjusting the speaking style across languages. Our experiments found that adding language information to the input of our multi-lingual autoregressive codec LM ϕ_{MAR} is surprisingly effective in guiding the right speaking style and relieving the L2 accent problem, which will be introduced in Section 5.5. Concretely, we embed language IDs into dense vectors and add them to the embeddings of acoustic tokens.

3.4 Cross-Lingual Inference

After training, VALL-E X can perform cross-lingual speech synthesis, as shown in Figure 3. In detail, we first concatenate source phonemes $\mathcal{S}^s$ and target phonemes $\mathcal{S}^t$ as prompts, and take the first-layer source acoustic tokens $\mathcal{A}^s_{:,1}$ as the decoding prefix, condition on which the multi-lingual autoregressive codec LM ϕ_{MAR} generates the first-layer target acoustic tokens $\mathcal{A}^t_{:,1}$,

$$\hat{a}^t_{i,1} \sim p_{\mathrm{AR}} \left(a^t_{i,1} \mid \left\langle \mathcal{S}^s, \mathcal{S}^t, \mathcal{A}^s_{:,1}, \mathcal{A}^t_{<i,1} \right\rangle ; \phi_{\mathrm{MAR}} \right), i = 1, \ldots, \tag{3}$$

where $\sim$ means probability-based sampling. The sampling is stopped until the `<end-of-sentence>` token is sampled. As mentioned in Section 3.3, language ID is used to control the speaking style of the final generated speech. After obtaining the first-layer target acoustic tokens $\mathcal{A}^t_{:,1}$ from ϕ_{MAR}, multi-lingual non-autoregressive codec LM ϕ_{MNAR} is used to predict the rest layers of acoustic tokens $\left\{ \mathcal{A}^t_{:,l} \mid l = 2, \ldots, 8 \right\}$ by greedy search, i.e., choosing the tokens with maximum probabilities,

$$\mathcal{A}^t_{:,l} = \underset{\mathcal{A}^t_{:,l}}{\operatorname{argmax}}\, p_{\mathrm{NAR}} \left(\mathcal{A}^t_{:,l} \mid \left\langle \mathcal{S}^t, \mathcal{A}^s_{:,1:8}, \mathcal{A}^t_{:,1:l-1} \right\rangle ; \phi_{\mathrm{MNAR}} \right), l = 2, \ldots, 8. \tag{4}$$

Finally, we use the decoder of EnCodec to synthesize the target speech from the complete target acoustic tokens $\mathcal{A}^t_{:,1:8}$.

[2] Current version of VALL-E X is trained on the speech-transcription of two languages, we leave exploring more languages for future work.

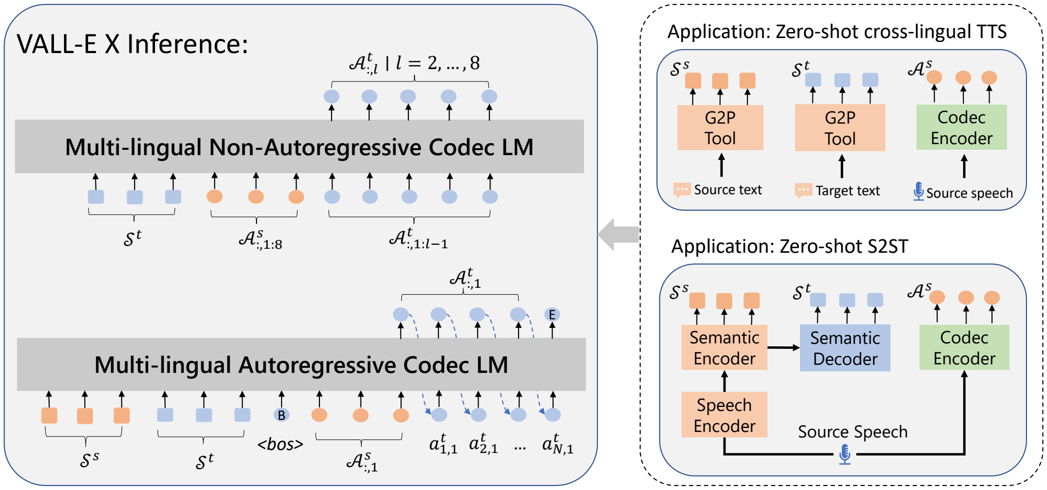

Figure 3: Inference illustration of the cross-lingual neural codec language model VALL-E X, with two-stage decoding strategies. VALL-E X can support zero-shot cross-lingual TTS and zero-shot speech-to-speech translation tasks.

4 VALL-E X Application

VALL-E X can be applied to various cross-lingual speech generation tasks. In this paper, we take zero-shot cross-lingual TTS and zero-shot speech-to-speech translation as two examples, as illustrated in Figure 3.

4.1 Zero-Shot Cross-Lingual TTS

The proposed VALL-E X is naturally suitable for zero-shot cross-lingual TTS tasks. Cross-lingual TTS tries to synthesize the target speech from text with a foreign speaker's voice. Conventional methods mainly employ additional speaker and language networks to model the speaker and language information respectively, without zero-shot synthesis capability. Thanks to the in-context learning capability of large language models, VALL-E X surprisingly shows the ability to perform zero-shot cross-lingual speech synthesis. More specifically, given the source speech, source transcript, and target text, we first convert source speech into source acoustic token $\mathcal{A}^s$ using the encoder of neural codec model EnCodec, and convert source transcript and target text into source phonemes $\mathcal{S}^s$ and target phonemes $\mathcal{S}^t$ using G2P tool. More specifically, as introduced in Section 3.4, we let $\mathcal{S}^t$ be the phonemes extracted from the target text, $\mathcal{S}^s$ and $\mathcal{A}^s$ be the phonemes and acoustic tokens extracted from the source speech. Then VALL-E X generates the full-layer target acoustic tokens, which are finally decompressed into the target speech by EnCodec decoder.

4.2 Zero-Shot Speech-to-Speech Translation

We can also apply our VALL-E X to zero-shot speech-to-speech translation tasks with additional speech recognition & translation model, which is responsible for synchronously recognizing and translating the source speech to the source and target phoneme sequences.

Speech Recognition & Translation Model We leverage the improved SpeechUT [Zhang et al., 2022c] as our speech recognition & translation model, which is a unified-modal speech-unit-text pre-training framework using hidden units as the modality bridge between speech and text. It supports various speech-to-text tasks, including both ASR and speech-to-text translation (ST). Inspired by SpeechLM [Zhang et al., 2022b] which explores different choices of units, we improve SpeechUT by replacing the clustering-based hidden units with phonemes. Specifically, it consists of a speech encoder, a phoneme encoder, and a phoneme decoder. All these components are pre-trained on ASR corpus (source speech $\mathcal{X}^{\mathrm{s}}$, source phoneme $\mathcal{S}^{\mathrm{s}}$) and MT corpus (source phoneme $\mathcal{S}^{\mathrm{s}}$, target phoneme $\mathcal{S}^{\mathrm{t}}$), where the phoneme sequences are converted from the text. Please see Appendix A.1.3 for more pre-training details about this model. After pre-training, the model is fine-tuned with ($\mathcal{X}^s$, $\mathcal{S}^s$, $\mathcal{S}^t$) triplet data derived from the ST corpus. Specifically, we perform multi-task learning with the CTC [Graves et al., 2006] loss added on the phoneme encoder predicting the source phonemes and the cross-entropy loss on the phoneme decoder predicting the target phonemes.

Inference Figure 3 shows the inference process of speech-to-speech translation. Given a source speech $\mathcal{X}^s$, the speech recognition & translation model first generates the source phonemes $\mathcal{S}^s$ from the semantic encoder and the target phonemes $\mathcal{S}^t$ from the semantic decoder. Besides, we use the EnCodec encoder to compress $\mathcal{X}^s$ into source acoustic tokens $\mathcal{A}^s$. Then, we concatenate $\mathcal{S}^s$, $\mathcal{S}^t$, and $\mathcal{A}^s$, as the input of VALL-E X, to produce the acoustic token sequence for the target speech, as introduced in Section 3.4. The generated acoustic tokens are converted to the final target speech with the decoder of EnCodec.

4.3 Evaluation

The proposed model is verified using various evaluation criteria, including speaker similarity (ASV-Score), speech quality (ASR-WER), translation quality (ASR-BLEU), naturalness, and human evaluation. Specifically, we measure speaker similarity between synthesized target speech and groud-truth target speech or source speech as an automatic speaker verification (ASV) task, where a WavLM [Chen et al., 2022] based ASV model is used to calculate the score. To verify the quality of generated speech, we first utilize the ASR system from the released HuBERT-Large model [Hsu et al., 2021] to recognize it into text. For TTS, speech quality is measured by ASR-WER between the recognized text and the original target text. For S2ST, speech quality is measured by ASR-BLEU between the recognized text and the provided translation text. Finally, to better verify our proposed VALL-E X systems, we adopt the open-source NISQA[3] [Mittag and Möller, 2021] (the NISQA-TTS model) to evaluate the naturalness of the synthetic speech and further perform the human evaluation with manual scoring on the generated speech, e.g., mean opinion score (MOS), comparative mean opinion score (CMOS) and similar mean opinion score (SMOS).

5 Experiments

We evaluate the proposed model on zero-shot cross-lingual TTS including English TTS prompted by Chinese speakers and Chinese TTS prompted by English speakers, and zero-shot S2ST including Chinese→English and English→Chinese directions. We provide the synthesized audio samples on our demo page to better show the performance of VALL-E X.

5.1 Dataset

Our VALL-E X is trained using bilingual speech-transcription (ASR) data. The Chinese ASR data are from WenetSpeech [Zhang et al., 2022a] containing 10,000+ hours of multi-domain labeled speech. The English ASR data are from LibriLight [Kahn et al., 2020] containing about 60,000 hours of unlabeled speech, whose speech data are collected from audiobooks. We train a Kaldi[4] ASR model on the labeled Librispeech [Panayotov et al., 2015] dataset to generate the pseudo transcripts for the unlabeled LibriLight speech.

To train the speech recognition & translation model for S2ST, we also use additional MT and ST data. The MT data are from AI Challenger[5], OpenSubtitles2018[6] and WMT2020[7], which contain about 13M, 10M, and 50M sentence pairs in conversion, drama[8], and news domains, respectively. The English→Chinese ST data is from GigaST [Ye et al., 2022], which is created by translating the transcripts in GigaSpeech [Chen et al., 2021] using a strong machine translation system. Similarly, we create the Chinese→English ST data by translating the transcripts of WenetSpeech using an MT model trained by ourselves on the MT data mentioned above.

[3] https://github.com/gabrielmittag/NISQA

[4] https://github.com/kaldi-asr/kaldi/tree/master/egs/librispeech

[5] https://challenger.ai/competition/translation

[6] https://opus.nlpl.eu/OpenSubtitles2018.php

[7] https://www.statmt.org/wmt20/translation-task.html

[8] http://www.opensubtitles.org/

We evaluate zero-shot S2ST using the Effective Multilingual Interaction in Mobile Environments (EMIME) dataset [Wester, 2010], which contains bilingual Chinese/English speech recorded by the same speakers. There are 25 pairs of bilingual sentences recorded by 7 female and 7 male native Chinese speakers, thus the total number of test examples is 350. Zero-shot cross-lingual TTS is evaluated using Librispeech [Panayotov et al., 2015] dev-clean set and EMIME dataset providing English and Chinese data, respectively. We have two settings in the experiments: (1) Librispeech English TTS with EMIME Chinese speech as prompts; (2) EMIME Chinese TTS with Librispeech Engish speech as prompts.

5.2 Experimental Setup

Phonemization & Quantization The right picture of Figure 2 illustrates the phonemization & quantization processes for different languages. All text data, including ASR transcripts and MT/ST translations, are converted by the lexicon provided in ASR datasets. We use a unified phoneme set called BigCiDian[9] for two languages which are based on International Phonetic Alphabet (IPA). The ASR transcripts (or pseudo transcripts) are also converted by Kaldi force-alignment tools[10] for additional alignment information used for the pre-training of speech recognition & translation model. The speech is quantized into discrete codec codes as the acoustic tokens using the neural audio codec model EnCodec[11], which employs residual vector quantization to iteratively quantize speech to a codebook according to the residual after quantization, resulting in multi-layer codebooks.

Model Architecture For the cross-lingual codec language models, ϕ_{MAR} and ϕ_{MNAR} are both 12-layer Transformer decoders with an attention dimension of 1024 and the FFN dimension of 4096. The autoregression is implemented by attention masking in the ϕ_{MAR} model. Sinuous position embedding is separately computed for each prompt sequence in ϕ_{MAR} and ϕ_{MNAR} models. Besides, the ϕ_{MNAR} model uses individual layer normalization for generating each layer of acoustic tokens. We also introduce the model architecture of speech recognition & translation for S2ST in Appendix A.1.2. We call our cross-lingual TTS model and S2ST model as **VALL-E X** and **VALL-E X Trans** in the subsequent experiments, respectively.

Training Details We optimize each module of VALL-E X individually, including ϕ_{MAR} and ϕ_{MNAR}. For both modules, The maximum sentence length is set to 20 seconds, so we re-segment the LibriLight data to an average utterance duration of 12 seconds by detecting the consecutive silence phonemes. Fortunately, the WenetSpeech data has already been segmented into short utterances. The maximum learning rate is 5e-4 with warm-up steps of 8,000. The models are trained on 32 V100 GPUs for 800k steps. ϕ_{MAR} is trained with the batch size of 120 seconds per GPU, which is 66 seconds for ϕ_{MNAR} due to the memory constraint. When optimizing ϕ_{MNAR}, instead of accumulating all layer's loss in Eqn. (2), we randomly select one layer at each optimization step for efficiency. For speech recognition & translation model, the training details can be found in Appendix A.1.3.

Baselines We adopt YourTTS[12] [Casanova et al., 2022] as our baseline for zero-shot cross-lingual TTS. YourTTS is a zero-shot multi-speaker TTS model for everyone, whose speaker information is based on speaker embedding extracted from a reference speech. Since previous work shows that current end-to-end S2ST systems underperform cascaded S2ST systems [Jia et al., 2022a, Lee et al., 2021b], we also build an S2ST baseline which is cascaded by an ASR model, an MT model, and a multi-speaker YourTTS. The source speech serves as the reference speech when synthesizing the target speech using YourTTS. The ASR model is the released HuBERT model introduced in Section 4.3, and the MT model is a vanilla Transformer trained by ourselves on the MT data introduced in Section 5.1. Since YourTTS is built only for English, we don't get its performance for English→Chinese translation direction.

[9] https://github.com/speechio/BigCiDian

[10] https://github.com/kaldi-asr/kaldi/tree/master/

[11] https://github.com/facebookresearch/encodec

[12] https://github.com/Edresson/YourTTS

Table 2: Zero-shot cross-lingual TTS evaluation for English TTS with Chinese speech as prompts and Chinese TTS with English speech as prompts, using automatic evaluation matrices, including ASV-Score (hypothesis vs. prompt), ASR-WER, and Naturalness.

	ASV-Score	ASR-WER	Naturalness
English TTS with Chinese as prompts			
Baseline (YourTTS)	0.30±0.10	8.53	3.36
VALL-E X	0.36±0.11	4.07	3.54
Chinese TTS with English as prompts			
VALL-E X	0.29±0.10	8.52	3.36

5.3 Zero-Shot Cross-Lingual TTS Evaluation

We first select samples with a length between 4 and 10 seconds from LibriSpeech dev-clean set, resulting in 40 speakers and 1373 samples. For English TTS, we randomly select one audio from EMIME set as the Chinese prompt for each target sentence in LibriSpeech dev-clean set. For Chinese TTS, we use extra 149 Chinese text sentences provided by the EMIME set and repeat them to the total number of 1373 so that they can be prompted by the LibriSpeech audios one-by-one. When synthesizing the target language speech, the whole sequence of the source language speech is used as the prompt.

Automatic Evaluation Table 2 summarizes the results of cross-lingual zero-shot TTS tasks, including English TTS prompted by Chinese speech and Chinese TTS prompted by English speech. We measure the speaker similarity using the automatic speaker verification (ASV) model, ranging from -1 to +1 given two speech utterances. The larger the value, the more similar the speakers of the two utterances are. The results show that: (1) for English TTS with Chinese as prompts, the speaker similarity between the hypothesis and prompts of VALL-E X is superior to that of the baseline (0.36 vs 0.30). (2) VALL-E X reduces the WER significantly from the baseline (from 8.53 to 4.07), demonstrating the effectiveness of our method. (3) VALL-E X has better speech naturalness than the baseline thanks to the large-scale training data and the large language model capacity. The results of Chinese TTS with English prompts are also listed.

Table 3: Human evaluation for zero-shot cross-lingual TTS. SMOS means similarity MOS between generated speech and prompt, and CMOS means comparative MOS based on Baseline.

	SMOS	CMOS (v.s. Baseline)
Baseline (YourTTS)	3.42±0.19	0.00
VALL-E X	4.00±0.20	+0.24

Human Evaluation We further conduct the human evaluation on 50 randomly selected speech records for zero-shot cross-lingual English TTS with Chinese speech as prompts, including SMOS and CMOS. Note that SMOS ranges from 1 to 5 where the larger the value, the higher the voice similarity, and CMOS ranges from -3 to 3 where the positive number means the new system is better than the baseline. The results are listed in Table 3. Baseline gets 3.42 SMOS scores between generated speech and prompts, while our VALL-E X achieves 4.00, which further demonstrates the model's superiority in keeping the speech characteristic in the cross-lingual setting. Moreover, to directly compare the speech synthesis quality between the proposed VALL-E X and baseline, we calculate the CMOS score between them evaluated by native speakers on the 50 sentences. The last column of Table 3 shows that VALL-E X obtains the gain of +0.24 CMOS scores than the baseline.

5.4 Zero-Shot S2ST Evaluation

S2ST is evaluated on bidirectional Chinese↔English data of EMIME dataset, measured by speaker similarity, translation quality, speech naturalness, and human evaluation.

Table 4: S2ST performance on EMIME dataset for Chinese↔English directions. Baseline is a cascaded S2ST system based on speaker embedding. Automatic evaluation matrices include ASV-Score, ASR-BLEU, and Naturalness.

	ASV-Score			ASR-BLEU	Naturalness
	tgt vs. src	hyp vs. src	hyp vs. tgt		
Chinese→English S2ST					
Baseline (S2ST)	0.58±0.09	0.28±0.10	0.27±0.12	27.49	3.44
- w/ oracle target text		0.28±0.10	0.29±0.11	80.30	3.43
VALL-E X Trans		0.37±0.10	0.37±0.11	30.66	3.54
- w/ oracle target text		0.39±0.10	0.38±0.10	86.78	3.54
English→Chinese S2ST					
VALL-E X Trans	0.58±0.09	0.48±0.11	0.53±0.11	34.45	3.41
- w/ oracle target text		0.47±0.12	0.55±0.11	84.00	3.42

Speaker Similarity We first evaluate whether the speaker's voice is preserved in the generated target speech using speaker similarity (ASV-Score), whose results are listed in Table 4. Because the EMIME test set has paired speech utterances with Chinese and English, we are able to calculate the ASV score among the generated speech (hyp), the source speech (src), as well as the target speech (tgt), resulting in 3 settings (tgt vs. src, hyp vs. src, and hyp vs. tgt). From Table 4 we can find that: (1) For Chinese→English, the ASV score of VALL-E X Trans significantly outperforms that of the conventional speaker embedding based S2ST system (Baseline), demonstrating the superiority of our model in terms of maintaining the source speaker's voice. (2) The ASV score has similar values when the generated speech (hyp) is compared with the source speech (src) and the target speech (tgt), and it is far away from the upper bound (tgt vs. src) for the English→Chinese direction, which suggests that the cross-lingual voice transferability still has the improvement space. (3) When directly generating speech from the ground-truth (oracle) text which degrades into cross-lingual TTS, the ASV score does not increase notably, indicating that voice transferability is less affected by the quality of translation.

Translation Quality Table 4 also shows the translation performance of VALL-E X Trans. Note that ASR-BLEU with oracle target text as the input of VALL-E X can be seen as the upper bound when translations are exactly correct. With oracle target text as input, VALL-E X Trans can achieve the performance of about 84∼87 BLEU scores, which also reflects the high performance of our neural codec language model. For Chinese→English, VALL-E X Trans achieves higher BLEU over the baseline (30.66 vs. 27.49), demonstrating the end-to-end speech-to-phoneme translation is more effective against the conventional cascaded speech-to-text translation when applying to S2ST task.

Speech Naturalness We also evaluate the Naturalness with the open-source NISQA [Mittag and Möller, 2021] for S2ST outputs. As shown in the last column of Table 4, compared to the baseline, VALL-E X Trans achieves a better naturalness score (3.54 vs. 3.44), which shows that VALL-E X can generate more natural target language speech than the baseline.

Human Evaluation We randomly sample 56 translation pairs[13] to perform a human evaluation using SMOS and MOS matrics for both Chinese→English and English→Chinese directions. Table 5 lists the results of VALL-E X Trans as well as the Chinese→English baseline. We use MOS (from 1 to 5 scores) instead of CMOS because the translated content may be different among models, which is not suitable for CMOS evaluation. For speaker similarity evaluation, VALL-E X Trans outperforms the baseline with 1.06 SMOS scores (4.12 vs. 3.06), demonstrating its superior ability to model speaker property of the proposed VALL-E X. Note that this value still can be improved since it is still far from the SMOS between the source speech prompt and ground truth (4.91). For speech quality, our VALL-E X slightly outperforms the baseline in Chinese→English S2ST in terms of MOS score (3.87 vs. 3.81).

[13] There are 14 speakers in the bilingual Chinese/English dataset, and 4 sentence pairs are chosen for each speaker, resulting in 56 translation pairs in total.

Table 5: Subjection evaluation with SMOS and MOS scores on bidirectional Chinese↔English S2ST tasks. SMOS is measured by comparing with the ground-truth target speech. English→Chinese S2ST baseline is not reported since it is not supported by the released YourTTS.

	Chinese→English		English→Chinese	
	SMOS	MOS	SMOS	MOS
Baseline (S2ST)	3.06±0.14	3.81±0.19	-	-
VALL-E X Trans	4.12±0.13	3.87±0.21	3.94±0.15	3.48±0.13
Source speech prompt	4.91±0.05	-	4.64±0.06	-
Oracle target speech	-	3.92±0.17	-	3.88±0.13

5.5 Analysis

In this section, we first analyze the effect of language ID, then explore the foreign accent problems, and qualitatively investigate the ability to maintain voice emotion and synthesize code-switch speech of our proposed model.

Effect of Language ID Our VALL-E X is trained with multi-lingual ASR data, which might increase the modeling difficulty for each specific language. We address it by adding language IDs to guide speech synthesis in the autoregressive language codec model. Here, we verify the effectiveness by removing the language ID (LID) or adding the wrong LID (i.e. the source LID). The ASV-Score and ASR-BLEU are reported in Table 6. Without LID or with the wrong language ID, the translation quality decreases, while the speaker similarity between the hypothesis and source speech increases. These results demonstrate the importance of language ID for the accuracy of the content. It also indicates that target LID reduces the transfer of information, which means the model without LID or with source LID will better maintain the sound of the original speaker.

Table 6: Evaluation for the effect of language ID on Chinese↔English EMIME dataset. ASV-Score is computed between synthesized speech and source prompt speech. The last column lists the subjection evaluation score of the foreign accent (from 1 to 5 scores).

	ASV-Score (vs. src)	ASR-BLEU	Accent Score
Chinese→English S2ST			
VALL-E X Trans	0.37±0.10	30.66	4.10
w/o Language ID	0.41±0.10	29.04	2.98
w/ wrong Language ID	0.41±0.10	29.07	2.55
English→Chinese S2ST			
VALL-E X Trans	0.48±0.11	34.45	4.03
w/o Language ID	0.49±0.11	30.86	2.35
w/ wrong Language ID	0.50±0.11	29.70	2.25

Foreign Accent Control L2 (second-language, or foreign) accent problem, the synthesized speech sounds like the accents of a foreigner, has arisen in cross-lingual TTS systems [Zhang et al., 2019, Lee et al., 2022]. Automatic Evaluation has shown that adding LID can boost speech quality. Besides, we conduct a subjection evaluation to label foreign accents from 1 to 5 on randomly selected 20 synthesized speech for both English and Chinese, where each sample is measured with a score from 1 to 5 denoting high-status foreign speakers, low-status foreign speakers, middle-status speakers, low-status native speakers, and high-status native speakers, respectively. As summarized in the last column of Table 6, we observed that our VALL-E X can control the accent for the target speech by LID modules. For example, in English→Chinese, VALL-E X Trans with right LID and without LID get the score of 4.03 and 2.35, respectively. This indicates that by using correct LID embedding, VALL-E X Trans is able to alleviate the foreign accent problem. Please also see the demo for audio examples of VALL-E X Trans with or without language ID.

Voice Emotion Maintenance Generating the speech with a specific emotion is a difficult task for speech synthesis because conventional TTS methods require TTS data with emotion labels to train

[Um et al., 2020]. Moreover, it is more tempting to reserve the source speaker's emotion in generated target speech for the S2ST task, which is not explored in previous S2ST work. In these experiments, we adopt the source prompts from the emotional voices dataset EmoV-DB [Um et al., 2020] as inputs of VALL-E X Trans to generate the translated target speech, whose samples are listed on our demo page. We found that the proposed VALL-E X can maintain emotional consistency to a certain extent between the source prompt and the synthesized speech. The underlying reasons are (1) our VALL-E X is trained with large-scale multi-lingual multi-speaker speech-transcription data, which contains various emotional speech records, and (2) the strong in-context learning ability of VALL-E X, like GPT-3 [Brown et al., 2020], promotes the generated speech to reserve the characteristic of the source prompt.

Code-Switch Speech Synthesis It is a common phenomenon to use code-switch utterances in bilingual or multi-lingual communities [Cao et al., 2020, Zhao et al., 2020, Manghat et al., 2022]. Code-switch speech synthesis aims to produce a fluent and consistent voice for code-switch text. Although our proposed VALL-E X is trained on multiple monolingual speech data, without special optimization for code-switch setting, VALL-E X provides a promising solution to code-switch speech synthesis. We put the code-switch samples on our demo page, demonstrating that due to its strong in-context learning ability, VALL-E X can synthesize fluent code-switch speech with a consistent voice.

6 Conclusion

In this work, we propose VALL-E X, a cross-lingual neural codec language model, which can retrain the source language speaker's voice in the generated target language speech. VALL-E X is free of the requirement for cross-lingual paired data from the same speakers. By training on large-scale multi-lingual multi-speaker speech-transcription data, the proposed VALL-E X demonstrates strong in-context learning capabilities and can support zero-shot cross-lingual text-to-speech and zero-shot voice-retentive speech-to-speech translation tasks. For future work, we plan to expand this method with more data and more languages.

References

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

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

Zexin Cai, Yaogen Yang, and Ming Li. Cross-lingual multi-speaker speech synthesis with limited bilingual training data. Computer Speech & Language, 77:101427, 2023.

Yuewen Cao, Songxiang Liu, Xixin Wu, Shiyin Kang, Peng Liu, Zhiyong Wu, Xunying Liu, Dan Su, Dong Yu, and Helen Meng. Code-switched speech synthesis using bilingual phonetic posteriorgram with only monolingual corpora. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7619–7623. IEEE, 2020.

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

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

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6): 1505–1518, 2022.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 244–250. IEEE, 2021.

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

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

Nikolaos Ellinas, Georgios Vamvoukakis, Konstantinos Markopoulos, Georgia Maniati, Panos Kakoulidis, June Sig Sung, Inchul Hwang, Spyros Raptis, Aimilios Chalamandaris, and Pirros Tsiakoulis. Cross-lingual text-to-speech with flow-based voice conversion for improved pronunciation. arXiv preprint arXiv:2210.17264, 2022.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning, ICML '06, page 369–376, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143891. URL https://doi.org/10.1145/1143844.1143891.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:3451–3460, 2021.

Rongjie Huang, Zhou Zhao, Jinglin Liu, Huadai Liu, Yi Ren, Lichao Zhang, and Jinzheng He. Transpeech: Speech-to-speech translation with bilateral perturbation. arXiv preprint arXiv:2205.12523, 2022.

Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. Direct speech-to-speech translation with a sequence-to-sequence model. arXiv preprint arXiv:1904.06037, 2019.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. Translatotron 2: Robust direct speech-to-speech translation. arXiv preprint arXiv:2107.08661, 2021.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In International Conference on Machine Learning, pages 10120–10134. PMLR, 2022a.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. Cvss corpus and massively multilingual speech-to-speech translation. arXiv preprint arXiv:2201.03713, 2022b.

J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 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, 2020. doi: 10.1109/ICASSP40776.2020.9052942.

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

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761, 2020b.

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

Alon Lavie, Alex Waibel, Lori Levin, Michael Finke, Donna Gates, Marsal Gavalda, Torsten Zeppenfeld, and Puming Zhan. Janus-iii: Speech-to-speech translation in multiple languages. In ICASSP, volume 1, pages 99–102. IEEE, 1997.

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

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. Textless speech-to-speech translation on real data. arXiv preprint arXiv:2112.08352, 2021b.

Jihwan Lee, Jae-Sung Bae, Seongkyu Mun, Heejin Choi, Joun Yeop Lee, Hoon-Young Cho, and Chanwoo Kim. An empirical study on l2 accents of cross-lingual text-to-speech systems via vowel space. arXiv preprint arXiv:2211.03078, 2022.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 6706–6713, 2019.

Xinjian Li, Ye Jia, and Chung-Cheng Chiu. Textless direct speech-to-speech translation with discrete speech representation. arXiv preprint arXiv:2211.00115, 2022.

Zhaoyu Liu and Brian Mak. Multi-lingual multi-speaker text-to-speech synthesis for voice cloning with online speaker enrollment. In INTERSPEECH, pages 2932–2936, 2020.

Sreeram Manghat, Sreeja Manghat, and Tanja Schultz. Normalization of code-switched text for speech synthesis. Proc. Interspeech 2022, pages 4297–4301, 2022.

Gabriel Mittag and Sebastian Möller. Deep learning based assessment of synthetic speech naturalness. arXiv preprint arXiv:2104.11673, 2021.

Eliya Nachmani and Lior Wolf. Unsupervised polyglot text-to-speech. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7055–7059. IEEE, 2019.

Satoshi Nakamura, Konstantin Markov, Hiromi Nakaiwa, Gen-ichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, J-S Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto. The atr multilingual speech-to-speech translation system. IEEE Transactions on Audio, Speech, and Language Processing, 14(2):365–376, 2006.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499, 2016.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 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, 2015. doi: 10.1109/ICASSP.2015.7178964.

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

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 464–468, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2074. URL https://aclanthology.org/N18-2074.

Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. Naturalspeech: End-to-end text to speech synthesis with human-level quality. arXiv preprint arXiv:2205.04421, 2022.

Se-Yun Um, Sangshin Oh, Kyungguen Byun, Inseon Jang, ChungHyun Ahn, and Hong-Goo Kang. Emotional speech synthesis with rich and granularized control. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7254–7258. IEEE, 2020.

Wolfgang Wahlster. Verbmobil: foundations of speech-to-speech translation. Springer Science & Business Media, 2013.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. arXiv preprint arXiv:2101.00390, 2021.

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

Kun Wei, Long Zhou, Ziqiang Zhang, Liping Chen, Shujie Liu, Lei He, Jinyu Li, and Furu Wei. Joint pre-training with speech and bilingual text for direct speech to speech translation. arXiv preprint arXiv:2210.17027, 2022.

Mirjam Wester. The emime bilingual database. Technical report, The University of Edinburgh, 2010.

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. arXiv preprint arXiv:2207.09983, 2022.

Jingzhou Yang and Lei He. Towards universal text-to-speech. In Interspeech, pages 3171–3175, 2020.

Jingzhou Yang and Lei He. Cross-lingual text-to-speech using multi-task learning and speaker classifier joint training. arXiv preprint arXiv:2201.08124, 2022.

Rong Ye, Chengqi Zhao, Tom Ko, Chutong Meng, Tao Wang, Mingxuan Wang, and Jun Cao. Gigast: A 10,000-hour pseudo speech translation corpus. arXiv preprint arXiv:2204.03939, 2022.

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

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6182–6186, 2022a. doi: 10.1109/ICASSP43922.2022.9746682.

Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, RJ Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. arXiv preprint arXiv:1907.04448, 2019.

Yu Zhang, Ron J Weiss, Byungha Chun, Yonghui Wu, Zhifeng Chen, Russell John Wyatt Skerry-Ryan, Ye Jia, Andrew M Rosenberg, and Bhuvana Ramabhadran. Multilingual speech synthesis and cross-language voice cloning, December 3 2020. US Patent App. 16/855,042.

Ziqiang Zhang, Sanyuan Chen, Long Zhou, Yu Wu, Shuo Ren, Shujie Liu, Zhuoyuan Yao, Xun Gong, Lirong Dai, Jinyu Li, et al. Speechlm: Enhanced speech pre-training with unpaired textual data. arXiv preprint arXiv:2209.15329, 2022b.

Ziqiang Zhang, Long Zhou, Junyi Ao, Shujie Liu, Lirong Dai, Jinyu Li, and Furu Wei. Speechut: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training. arXiv preprint arXiv:2210.03730, 2022c.

Shengkui Zhao, Trung Hieu Nguyen, Hao Wang, and Bin Ma. Towards natural bilingual and code-switched speech synthesis based on mix of monolingual recordings and cross-lingual voice conversion. arXiv preprint arXiv:2010.08136, 2020.

A Appendix

A.1 Speech Recognition & Translation Model

A.1.1 Model Pre-training

Specifically, speech recognition & translation model consists of a speech encoder (θ_{enc1}), a semantic encoder (θ_{enc2}), and a semantic decoder (θ_{dec}). Given a speech waveform $\mathcal{X}^s$ and the corresponding phonemes $\mathcal{S}^s \triangleq \{s_i^s | i = 1, \ldots, N\}$ where N is the sequence length, the speech-side pre-training objective is to predict the phonemes from the top of the speech encoder and semantic encoder, formalized as

$$\mathcal{L}_{\text{speech}} = -\sum_{i \in \mathcal{M}} \left(\log p\left(s_i^s | \mathcal{X}^s; \theta_{enc1}\right) + \log p\left(s_i^s | \mathcal{X}^s; \theta_{enc1}, \theta_{enc2}\right)\right) \tag{5}$$

where $\mathcal{M}$ is a set of masked positions, and the $p(.)$ is parameterized as the same way with original SpeechUT. Then, given bilingual phoneme sequences, $\mathcal{S}^s$ and $\mathcal{S}^t$, the text-side pre-training objective is to perform sequence-to-sequence translation autoregressively, formalized as

$$\mathcal{L}_{\text{text}} = -\sum_{i=1}^{|\mathcal{S}^t|} \log p\left(s_i^t | \mathcal{S}_{<i}^t, \mathcal{S}^s; \theta_{enc2}, \theta_{dec}\right) \tag{6}$$

In this way, each of the three components can be pre-trained with one or two learning objectives. The final pre-training objective is $\mathcal{L}_{\text{pt}} = \mathcal{L}_{\text{speech}} + \mathcal{L}_{\text{text}}$.

A.1.2 Model Architecture

For the speech recognition & translation model, we leverage the Base architecture of the SpeechUT model, where all encoder/decoders consist of 6 Transformer layers with relative position bias [Shaw et al., 2018]. The FFN dimension is 3072 and the attention dimension is 768. Besides, a speech pre-net is equipped before the speech encoder, which contains several 1-D convolutional layers with 512 channels and kernel sizes of [10,3,3,3,3,2,2]. It can downsample the speech waveform by 320 and convert it to fix-dimensional embeddings.

A.1.3 Training Details

The speech recognition & translation model is pre-trained following the hyper-parameter setting of Zhang et al. [2022c]. The speech mask probability is 8% and the mask length is 10. The embedding mixing mechanism of the original SpeechUT is also performed. The batch sizes of speech and phonemes on each GPU are 1,400,000 (87.5 seconds) and 3,000, respectively. The maximum learning rate is 5e-4 with warm-up steps of 32,000. The model is pre-trained on 32 V100 GPUs for 400K steps. After pre-training, we perform ASR/ST joint fine-tuning, where the transcription phonemes are predicted on the top of the semantic encoder through a nonlinear CTC layer, and the translation phonemes are predicted through the semantic decoder. the transcription phonemes are reduced by removing the repetitive phonemes. During fine-tuning, we empirically set the weight of the CTC loss to 0.2. The models are tuned on 32 GPUs with a batch size of 2,000,000 (125 seconds) per GPU for 200K steps.