POLYVOICE: LANGUAGE MODELS FOR SPEECH TO SPEECH TRANSLATION

Qianqian Dong*, **Zhiying Huang***, **Qiao Tian, Chen Xu, Tom Ko**†,
Yunlong Zhao, Siyuan Feng, Tang Li, Kexin Wang, Xuxin Cheng, Fengpeng Yue,
Ye Bai, Xi Chen, Lu Lu, Zejun Ma, Yuping Wang, Mingxuan Wang, Yuxuan Wang
ByteDance
{dongqianqian, huangzhiying.92, tom.ko}@bytedance.com

ABSTRACT

With the huge success of GPT models in natural language processing, there is a growing interest in applying language modeling approaches to speech tasks. Currently, the dominant architecture in speech-to-speech translation (S2ST) remains the encoder-decoder paradigm, creating a need to investigate the impact of language modeling approaches in this area. In this study, we introduce PolyVoice, a language model-based framework designed for S2ST systems. Our framework comprises three decoder-only language models: a translation language model, a duration language model, and a speech synthesis language model. These language models employ different types of prompts to extract learned information effectively. By utilizing unsupervised semantic units, our framework can transfer semantic information across these models, making it applicable even to unwritten languages. We evaluate our system on Chinese $\rightarrow$ English and English $\rightarrow$ Spanish language pairs. Experimental results demonstrate that PolyVoice outperforms the state-of-the-art encoder-decoder model, producing voice-cloned speech with high translation and audio quality. Speech samples are available at https://polyvoice.github.io.

1 INTRODUCTION

Speech-to-speech translation (S2ST) is a challenging task as it encounters all the difficulties of automatic speech recognition (ASR), machine translation (MT) and text-to-speech (TTS) synthesis. The research in S2ST focuses on two approaches: cascade solutions (Lavie et al., 1997; Baldridge, 2004; Nakamura et al., 2006) and direct solutions (Jia et al., 2019; 2022a). The advantage of cascade solutions lies in the convenience of improving the performance of individual modules, while direct solutions excel in lower latency and simpler model architectures. As for the direct S2ST solutions, it used to involve direct output of mel-spectrogram features (Dong et al., 2022) in the early stages, but recently, there has been a growing interest in predicting discrete units (Lee et al., 2022a). The unit-based approach has become more popular due to several reasons: (1) It eases the modeling difficulty of emitting spectrogram. (2) Units can be generated through unsupervised methods and can cover unwritten languages. (3) It allows a connection with other token-based NLP models.

Recently, language modeling (LM) approaches have made a lot of breakthroughs in natural language processing (NLP) (Zhao et al., 2023). The success of GPT models (Brown et al., 2020; Ouyang et al., 2022) is leading the community to a new era. Currently, the encoder-decoder models remain widely used in speech tasks, and the exploration of using LM approaches is still in its early stages. In fact, there have been attempts on ASR (Fathullah et al., 2023) and TTS (Wang et al., 2023), indicating that this direction is promising. Thus, we are motivated to investigate the performance of language modeling approach in S2ST. In this paper, we propose a semantic unit-based framework for S2ST system. Our framework (Fig. 1) consists of three LMs: a translation LM, a duration LM and a speech synthesis LM. The translation LM processes the semantic units of the source language and translates them into semantic units of the target language. The duration LM predicts the duration

*Equal contribution.
†Corresponding author.

information of the target semantic units and extends the unit sequence. The speech synthesis LM predicts the target acoustic units which are then converted into a waveform by a unit vocoder.

We employ various prompt types to extract the acquired knowledge from the language models utilized in our approach. Specifically, we concatenate the source and target semantic units, along with the source acoustic units, forming a comprehensive prompt for the speech synthesis language model. This enables the speech synthesis language model to grasp the acoustic characteristics of the source speaker and generate voice-cloned acoustic units accordingly. Importantly, both the semantic and acoustic units mentioned above are generated through unsupervised methods, making our framework applicable to unwritten languages.

We evaluate our system on Chinese → English and English → Spanish language pairs. Experimental results show that our system can generate voice-cloned speech with high translation quality and audio quality. We summarise our contribution as follows:

- We propose using a series of decoder-only language models to fulfill the S2ST task, whereas encoder-decoder models are the dominant structure in previous works.
- Unsupervised speech units are used in the framework and thus PolyVoice can cover both written and unwritten languages.

The rest of this paper is organized as follows. Section 2 introduces related work. Details of our method are described in Section 3. Section 4 introduces our experimental setup and main results. Section 5 presents our ablation study. Finally, we conclude our work in the last section.

2 Related Work

2.1 Speech Tokenization

There are two kinds of discretized speech units used in our work: semantic and acoustic units. Semantic units are usually derived from representations produced by speech encoder models like HuBERT (Hsu et al., 2021), mHuBERT (Lee et al., 2022c) or w2v-BERT (Chung et al., 2021). They capture the phonetics and semantic content in speech. Although the making of these units is originally developed to be used as target for training the speech encoder, recently there are attempts to directly use these units as input/output for semantic tasks (Kharitonov et al., 2021; Lakhotia et al., 2021; Meng et al., 2023; Zhang et al., 2023a). Acoustic units can also be referred to as codec units. They are originally developed to transmit high-quality speech signals under limited bandwidth. AudioLM (Borsos et al., 2023) is a pioneer work in using language models (LM) for audio generation. They make use of both kinds of units and build several LMs with different resolutions. VALL-E (Wang et al., 2023) further extends the AudioLM framework and applies it in TTS. They successfully demonstrate that the in-context learning capabilities of LM can be similarly replicated in the context of phoneme and codec units. In contrast to phoneme units which have to involve a supervised training process, both semantic and acoustic units can be generated through unsupervised methods.

2.2 TTS

Zero-shot cross-lingual TTS (Jia et al., 2018; Cooper et al., 2020) aims to build a system that can synthesize speech with user's voice and in a specific language that the user doesn't speak. Early attempts include speaker adaptation (Chen et al., 2019) and speaker embedding (Liu & Mak, 2019) approaches. LM-based TTS has been recently proposed and demonstrated promising results. VALL-E (Wang et al., 2023) introduces an LM-based approach that leverages in-context learning for zero-shot TTS. Their approach utilizes phoneme units and source acoustic units to prompt the LM in predicting the target acoustic units. The most relevant related work to ours is VALL-E X (Zhang et al., 2023b), which extends the VALL-E framework to tackle the cross-lingual problem. In their work, they concatenate the source and target phoneme units, along with the source acoustic units, to create a prompt for the LM.

2.3 S2ST

Speech-to-speech translation (Lavie et al., 1997; Baldridge, 2004; Nakamura et al., 2006) aims to develop models capable of generating target language speech from source language speech. A vanilla system traditionally employs a pipeline (Nakamura et al., 2006) that sequentially processes the input through automatic speech recognition (ASR) models, machine translation (MT) models, and text-to-speech synthesis (TTS) models. Recently, end-to-end paradigms (Jia et al., 2019; 2022a) have gained popularity in the field of S2ST, as they allow for a single model to perform one or more of the aforementioned tasks, which consequently reduces error propagation and latency. Among the various techniques, auxiliary supervision based on textual data has been particularly effective during training (Jia et al., 2019; Kano et al., 2021). However, this approach is not feasible when dealing with unwritten languages. To address this challenge, discrete units (Hsu et al., 2021) extracted from the speech are used to replace the target text, and then can be synthesized into the speech (Tjandra et al., 2019; Zhang et al., 2021; Lee et al., 2022a). Large scale studies have shown the powerful performance in various speech processing tasks (Nguyen et al., 2022).

Current research in speech-to-speech translation primarily emphasizes translation quality, with notable improvements observed in automatic evaluation metrics (like BLEU) or human evaluation of naturalness. However, there remain two persistent challenges in developing practical systems. First, these systems are predominantly developed and evaluated on small-scale benchmarks, while real-world scenarios often involve large quantities of labeled data, including ASR, MT, and S2T data (Agrawal et al., 2023). Even for low-resource or unwritten languages, leveraging unlabeled speech or text can provide valuable information (Lee et al., 2022a). Therefore, developing a unified model that jointly utilizes various data types is a critical research goal yet to be achieved. Second, while not a strict requirement, preserving the source speaker's style during translation is an important aspect of improving user experience (Zhang et al., 2023b; Song et al., 2023). However, capturing the unique characteristics of individual speakers is a challenging task. Current approaches, such as speaker embeddings (Jia et al., 2019) and multi-speaker TTS systems (Jia et al., 2018), have made some progress in this direction, but they are still far from practical requirements.

Taking a broad perspective, our work aligns with the framework presented in Lee et al. (2022a), where the source speech undergoes translation into discrete units before synthesizing it into the target language's speech. However, what distinguishes our work is the utilization of decoder-only language models to enhance the performance of each module. By leveraging diverse data sources within a language model-based framework, our proposed method effectively maintains the source speaker's style during synthesis, thereby demonstrating significant potential in practical systems.

3 METHOD

We present PolyVoice, an innovative language model-based framework for speech-to-speech translation, catering to both written and unwritten languages. Our proposed framework leverages discrete units, acquired through self-supervised training techniques such as HuBERT (Hsu et al., 2021), serving as an intermediary representation bridging the source and target speech modalities. PolyVoice provides a comprehensive framework consisting of two main components: a speech-to-unit translation (S2UT) front-end, facilitating the conversion of source language speech into target language units, and a unit-to-speech (U2S) back-end, skillfully synthesizing translated speech while preserving the personalized style of the source speaker. Figure 1 provides an illustrative overview of our approach.

3.1 SPEECH-TO-UNIT TRANSLATION (S2UT)

By employing discrete units obtained through self-supervised training, semantically irrelevant information from continuous speech representations is eliminated, facilitating effective training in an NLP paradigm. In this regard, the S2UT component leverages a language model to acquire the necessary cross-lingual generation capabilities based on the unit-based approach.

Semantic unit extractor S2UT initiates the processing of raw speech data by employing a sophisticated semantic unit extractor. Here we adopt HuBERT, which first encodes the speech by a stack of convolutions and Transformer layers to continuous representations at every 20-ms frame,

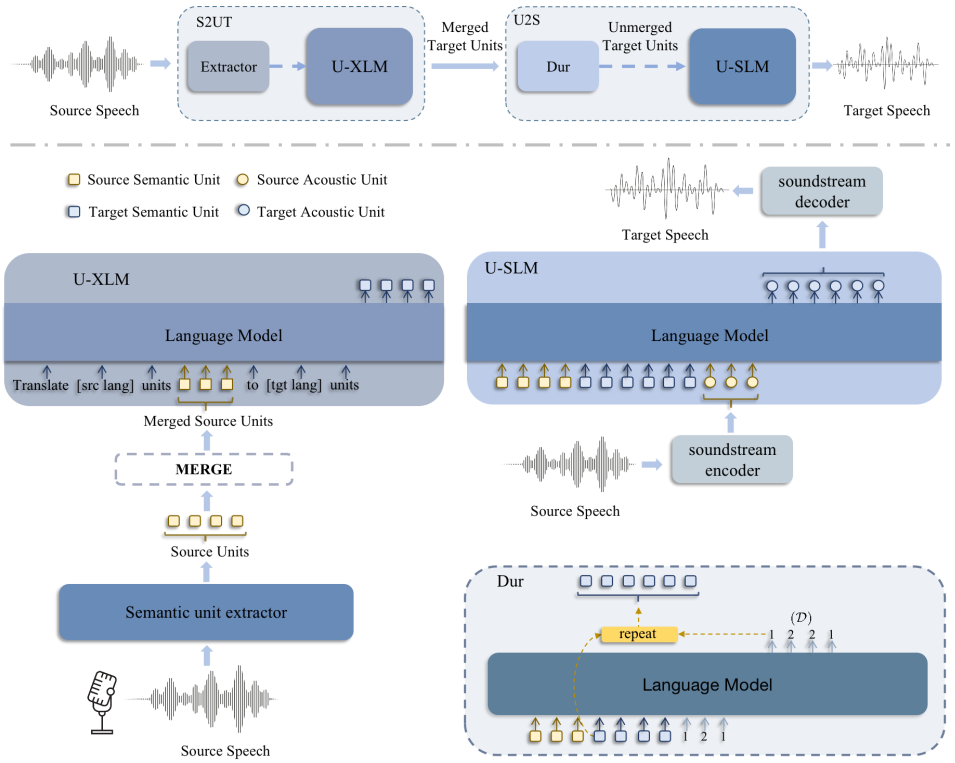

Figure 1: Overview of PolyVoice. The framework can be viewed as a concatenation of a speech-to-unit translation (S2UT) front-end and a unit-to-speech (U2S) back-end. There are three LMs in the framework: the unit-based cross-lingual LM (U-XLM), the duration LM and the unit-to-speech LM (U-SLM).

ASR: [lang]
Data: <**unit, text**>
Prompt1: *Translate **[lang]** unit " **{unit}** " to **[lang]** text: " **{text}** "* **Prompt2**: *Translate **[lang]** text " **{text}** " to **[lang]** unit: " **{unit}** "*
MT: [src_lang] → [tgt_lang]
Data: <**src_text, tgt_text**>
Prompt: *Translate **[src_lang]** text " **{src_text}** " to **[tgt_lang]** text: " **{tgt_text}** "*
S2ST: [src_lang] → [tgt_lang]
Data: <**src_unit, tgt_unit, src_text, tgt_text**>
Prompt1: *Translate **[src_lang]** unit " **{src_unit}** " to **[tgt_lang]** unit: " **{tgt_unit}** "* **Prompt2**: *Translate **[src_lang]** unit " **{src_unit}** " to **[src_lang]** text: " **{src_text}** "* **Prompt3**: *Translate **[src_lang]** unit " **{src_unit}** " to **[tgt_lang]** text: " **{tgt_text}** "* **Prompt4**: *Translate **[src_lang]** text " **{src_text}** " to **[tgt_lang]** unit: " **{tgt_unit}** "* **Prompt5**: *Translate **[src_lang]** text " **{src_text}** " to **[tgt_lang]** text: " **{tgt_text}** "* **Prompt6**: *Translate **[tgt_lang]** text " **{tgt_text}** " to **[tgt_lang]** unit: " **{tgt_unit}** "* **Prompt6**: *Translate **[src_lang]** unit " **{src_unit}** " to **[src_lang]** text then **[tgt_lang]** text then **[tgt_lang]** unit: "**{src_text}** [sep] **{tgt_text}** [sep] **{tgt_unit}** "* ...

Table 1: Data construction for the U-XLM model by various templates. These templates play a crucial role in generating multiple versions of training samples from diverse data resources such as ASR, MT, and S2ST, which are instrumental in facilitating cross-lingual unit generation.

and then utilizes k-means clustering to discretize the representation to a set of cluster indices $Z = z\_1, \cdots, z\_T$. T is the number of frames and $z\_t \in [K]$, where K is the number of cluster centroids. The discretized units are then merged by removing consecutive duplicated units.

Unit-based cross-lingual language model (U-XLM) We denote the training sample consisting of units of speech in source language and target language as <*src_unit, tgt_unit*>. Within the encoder-decoder architecture, the encoder takes the source units as input, while the decoder generates predictions for the target units. To facilitate the generation of cross-lingual units, a straightforward approach involves utilizing simple prompts to construct training samples for natural language from unit pairs. For instance, one can create prompts like: "*Translate* ***[src_lang]*** *unit* **{*src_unit*}** *to* ***[tgt_lang]*** *unit:* **{*tgt_unit*}**". In addition to the direct transformation prompt, we can also instruct the model to generate intermediate steps similar to the cascaded systems in the *chain of thought prompting* (Wei et al., 2022) manner.

Training To achieve competitive performance in training the U-XLM model, a large amount of data is crucial. However, obtaining supervised data, specifically cross-lingual unit pairs, is often limited in real-world scenarios. While auxiliary models can generate pseudo labels, such as synthesizing the target speech using the TTS model, direct training with supervised data is preferred.

To overcome the challenge of limited data availability, prior research has incorporated additional loss functions into the encoder-decoder architecture using multitask learning (Jia et al., 2022a; Lee et al., 2022a). In our work, we leverage the power of language modeling to address this issue in a more straightforward manner, allowing for the utilization of diverse data sources like automatic speech recognition (ASR) and machine translation (MT) data.

In Table 1, we demonstrate how we slightly modify the prompts to create training samples for different types of data sources. By employing parameter sharing and simplifying the design of auxiliary objectives, we train the model using these modified prompts. This approach also enables the direct utilization of unlabeled text and speech data. Consequently, the model implicitly enhances the alignment of the representation space between speech units and text.

3.2 Unit-to-speech Synthesis (U2S)

Unit-to-speech language model (U-SLM) As illustrated in Figure 1, the U-SLM leverages the semantic units predicted by U-XLM and generates the codec units which incorporate the speaking style of source speaker. Similar to VALL-E X, U-SLM encompasses both an autoregressive model and a non-autoregressive model. However, instead of conventional phonemes, we employ discretized semantic units in our approach.

SoundStream codec We employ SoundStream (Zeghidour et al., 2021), a neural audio codec, to generate embeddings of acoustic tokens. To ensure optimal performance, we re-implement the SoundStream with a hierarchy of 6 vector quantizers and a vocabulary of 1024 symbols. In our configuration, the acoustic tokens are produced at a rate of 80Hz for input waveforms sampled at 24 kHz, which results in a significant reduction in the sampling rate, specifically a reduction by a factor of 300 ($24000/80$). Once the U2S model predicts the acoustic tokens represented by the SoundStream codec, the decoder component of SoundStream reconstructs them back into the waveform.

Duration model Through empirical observations, we have determined that the duration information of discretized units is crucial for ensuring stable and natural-sounding synthesized speech. In our approach, we employ an additional LM to predict the durations.

In Figure 1, we illustrate the process of incorporating duration prediction into our framework. The merged source semantic unit sequence, merged target semantic unit sequence, and the source duration value sequence ($\mathcal{D}$) are concatenated and provided as a prompt to the duration LM. Subsequently, the duration LM predicts the target duration value sequence, and each target semantic unit is repeated accordingly based on its predicted duration.

4 Experiments

We conduct experiments utilizing a decoder-only model architecture following the standard GPT-2 (Radford et al., 2019). In Appendix A.2, we provide a comprehensive description of the model configurations employed in our work.

4.1 DATASETS AND PREPROCESSING

4.1.1 S2UT

Semantic tokens U-XLM is trained by cross-lingual unit data, which is extracted from the audio by HuBERT (Hsu et al., 2021) models. For Chinese audio, we utilize an open-source model based on WenetSpeech Chinese speech[1]. For English and Spanish audio, we use an open-source multilingual model (English, Spanish and French)[2]. The cluster centroids of a k-mean algorithm for the two models are 500 and 1,000, respectively.

Vocabulary To address the out-of-vocabulary problem and enable parameter sharing across languages, we utilize byte-level subword units[3] that decompose each character into byte-sized pieces and achieve a final vocabulary size of 56,407, including 1,500 cluster centroids (<zh-0>,...<zh-499> and <m-0>,...,<m-999>).

Datasets Considering that the paired speech-to-speech (S2S) data is scarce, we synthesize the pseudo data from the ASR data utilizing in-house MT and TTS systems. In addition, various types of data resources provide better learning of the U-XLM model, like large-scale ASR and MT data. A more elaborate description of the used datasets can be found in Appendix Table 7.

The S2S data is sourced from WenetSpeech (Zhang et al., 2022) and GigaSpeech (Chen et al., 2021), respectively, marked as "GigaS2S" and "WenetS2S". WenetSpeech is a Chinese ASR dataset with over 10,000 hours of speech data collected from YouTube. And we utilize a subset of 10,000 hours of GigaSpeech (Chen et al., 2021), an English ASR dataset collected from audiobooks, podcasts, and YouTube.

Then we scale up the training data using specific prompts for various types of datasets. We utilize the LibriLight (Kahn et al., 2020) and the in-house ASR datasets. LibriLight is an unlabeled English speech dataset containing about 60,000 hours of speech. Since LibriLight has many long audios, we segment and recognize the audio based on the method of voice active detection (VAD) and in-house ASR system, generating the audio length ranging from 0.5 to 25s, and the average length is 7s. The in-house ASR dataset is a Chinese ASR dataset with 60,000 hours of speech. We also use the in-house Chinese-English MT dataset consisting of 44M sentence pairs.

4.1.2 U2S

The U-SLM is trained on the large open-source bilingual speech data, including WenetSpeech (Zhang et al., 2022) and LibriLight (Kahn et al., 2020). The Librilight is handled in the same way as U-XLM. WenetSpeech keeps the original data length unchanged. The duration of audio samples ranges from 0.5 to 20s, and the average duration is 2.5s. To further improve the synthesized quality, we use an additional 250-hour internal Chinese TTS data and 400-hour internal English TTS data.

4.2 EVALUATION

Our method is evaluated on two speech-to-speech benchmark datasets, EMIME (Wester & Liang, 2011) (Chinese $\rightarrow$ English) and CVSS (Jia et al., 2022b) (English $\rightarrow$ Spanish). Apart from the overall result, we report the separate performance on the S2UT front-end and U2S back-end. EMIME contains bilingual Chinese-English speech recorded by the same speakers. For CVSS, the translation speech is in voices automatically transferred from the corresponding source speech. To measure the performance of our system, we evaluate both the translation quality and the speech quality.

Translation Quality Following the previous setups, we recognize the speech output by an in-house ASR system to compute BLEU scores (ASR-BLEU) for S2ST results using `sacrebleu`[4].

[1] https://github.com/TencentGameMate/chinese_speech_pretrain

[2] https://github.com/facebookresearch/fairseq/blob/main/examples/speech_to_speech/docs/textless_s2st_real_data.md

[3] https://github.com/huggingface/tokenizers

[4] https://github.com/mjpost/sacrebleu

	ASV ↑			ASR-BLEU ↑	Naturalness ↑
	tgt vs. src	hyp vs. src	hyp vs. tgt		
Cascade (VALL-E X paper)	0.58	0.28	0.27	27.49	3.44
+ w/ oracle target text		0.28	0.29	80.30	3.43
VALL-E X (VALL-E X paper)		0.37	0.37	30.66	3.54
+ w/ oracle target text		0.39	0.38	**86.78**	3.54
S2UT	0.59	0.06	0.08	29.30	3.35
PolyVoice (S2UT + U2S)		0.38	0.38	29.40	4.10
+ w/ chain of thought decoding		**0.38**	**0.38**	**30.80**	**4.11**
+ w/ oracle target semantic unit		**0.42**	**0.48**	76.10	**3.92**

Table 2: S2ST results on Chinese-English EMIME dataset. The **bold** and underlined numbers represent the best results of full-pipeline decoding. Only **bold** numbers signify the best synthesized results using the oracle target translation as input.

Speech Quality The speech quality is evaluated by multiple metrics. The capability of voice clone is measured by the speaker similarity (ASV-Score), which is calculated by an ASV model[5] to determine whether the synthesized speech is from the same speaker as the ground-truth speech. The naturalness of the speech output is evaluated by the automatic metric using `NISQA`[6]. And the pronunciation accuracy is evaluated using WER scores (ASR-WER) with an ASR model based on `hubert-large`[7].

4.3 Results and Analysis

4.3.1 S2ST Results

Table 2 summarizes the overall performance of our method for S2ST. We conduct experiments on the EMIME dataset to enable direct comparisons with the most similar work VALL-E X. The cascade system treats S2ST as a pipeline of running an ASR model, an MT model, and a multi-speaker YourTTS model separately and sequentially. During the synthesis process, speaker information is integrated using speaker embeddings.

We first evaluate the capability to preserve the voice of the source speaker in the output speech, using the ASV score. We calculate speaker similarity between the source speech, target speech, and synthesized speech. We can run the U-XLM alone, where speech is synthesized by a Unit-based vocoder[8] (Lee et al., 2022c). Due to the lack of explicit modeling of speaker characteristics, it produces particularly low ASV scores. Both the VALL-E X and PolyVoice systems, which adopt in-context learning, show superior performance over the speaker embedding based method. Notably, our method demonstrates better voice cloning capabilities when ground-truth target information is available.

PolyVoice achieves a slightly enhanced translation quality (ASR-BLEU) but a remarkable improvement in speech quality (naturalness) compared with VALL-E X. When taking the ground-truth target information as input, PolyVoice is inferior to VALL-E X with a large gap of about 10 BLEU points, while the naturalness improves significantly. The semantic units are extracted from the speech by unsupervised learning, which inevitably introduces errors. Although units are considered "semantic" tokens, they still preserve some acoustic information. Therefore, unit-based modeling leads to better speech quality but worse translation quality. In contrast, phonemes obtained from the text ensure semantic correctness but lose the acoustic information. And future work can focus on enhancing the extraction of semantic information to improve translation quality.

[5] https://github.com/Sanyuan-Chen/UniSpeech/tree/t-schen/asv_eval/downstreams/speaker_verification#example-2

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

[7] https://huggingface.co/facebook/hubert-large-ls960-ft

[8] https://github.com/facebookresearch/fairseq/blob/main/examples/speech_to_speech/docs/textless_s2st_real_data.md

CVSS	ASV ↑	BLEU ↑	Naturalness↑
Ground-truth	0.19	89.3	3.54
PolyVoice	0.34	18.3	3.60
+ w/ oracle target unit	0.28	70.8	3.69

Table 3: Results on the English-Spanish CVSS dataset. We train the model with paired speech-to-speech datasets expanded from GigaSpeech without any text information. BLEU means ASR-BLEU, target unit means oracle Spanish unit.

Arch	Training Data	ASR-BLEU
Encoder-Decoder	GigaS2S WenetS2S	16.8
+ w/ U2S		18.7
Decoder-only		20.7 (+3.9)
+ w/ U2S		22.0 (+3.3)

Table 4: Performance with different architectures on EMIME dataset.

Task	S2ST (BLEU ↑)	ASR (CER ↓)	ST (BLEU ↑)	MT (BLEU ↑)	TTS (WER ↓)
S2S	22.2	-	-	-	-
+ MTL	29.4	4.46	30.8	33.81	6.99

Table 5: The performance of multiple tasks on EMIME dataset. Here are the explanations for each task. S2ST: Chinese speech to English speech; ASR: Chinese speech to Chinese text; ST: Chinese speech to English text; MT: Chinese text to English text; TTS: English text to English speech.

Interestingly, PolyVoice achieves better naturalness using the predicted units. We speculate that this is due to the language model's output having better fluency. U-XLM learns the speech distribution over the large scale of unit data, and tends to generate more natural sequences of units. However, this may interfere with the accuracy of the translation. We will explore this issue in the future.

4.3.2 Unwritten Language Scenario

We examine our proposed framework in the case where the source is a written language and the target is an unwritten language. In our setup, we train and evaluate an English→Spanish S2ST system without the use of any Spanish text transcript. Table 3 summarizes the results. The ASR-BLEU (18.3) indicates that the Spanish speech generated by our system is semantically understandable. This demonstrates the ability of our S2ST system for the unwritten languages.

5 Ablation Study

5.1 Decoder-only vs. Encoder-Decoder

Empirical studies in the field of natural language processing have revealed that the full potential of the decoder-only approach can be realized through the use of large model sizes and expansive datasets. As pioneers in exploring the application of language models to S2ST, we present a fair comparison of the two architectures, *decoder-only* and *encoder-decoder*, in Table 4. Two frameworks are trained with the same training data and similar parameters, approximately 0.6B in size. Interestingly, the decoder-only model yields a remarkable improvement of 3.9 BLEU points over the encoder-decoder counterpart[9] (Lee et al., 2022b). When we synthesize the speech by U2S instead of vocoder, the performance gap is reduced, highlighting the robustness of our U2S back-end.

[9]The encoder-decoder architecture is experimented using the implementation: https://github.com/facebookresearch/fairseq/blob/main/examples/speech_to_speech/docs/direct_s2st_discrete_units.md.

Methods	WER ↓	ASV ↑	Naturalness ↑
VALL-E X (paper)	4.07	0.36	3.54
U2S	6.40	0.38	3.98
+ w/o semantic2dur	31.93	0.37	3.81
+ w/ mHuBERT_zh_en	4.76	0.37	3.81

Table 6: Evaluation of the speech synthesizers.

5.2 Multi-task Training

As discussed in Section 3, the language modeling enables direct training over the diverse data sources utilizing specific prompts. In this way, we combine additional large-scale ASR and MT data to fully explore the potential of our method. As shown in Table 5, U-XLM achieves promising performance for multiple tasks involved (including S2ST, ASR, ST, MT, and TTS) under the expanded data setting, which verifies the capability of the general modeling in the decoder-only architecture. In the traditional paradigm, we need to design a complex manner to combine multi-task learning, but language modeling only modifies the prompt to construct the training data.

5.3 Zero-shot Cross-lingual Unit-to-speech

We select samples with a duration between 4 and 10 seconds from LibriSpeech (Panayotov et al., 2015) `dev-clean` set to evaluate the zero-shot cross-lingual unit-to-speech module. And we randomly choose one audio from EMIME as the Chinese speech prompt.

Table 6 shows the resynthesis performance of different speech synthesizers. Our TTS obtains better performance in both ASV and naturalness. We attribute the increase of WER to the difference in the amount of semantic information carried by phonemes and unsupervised units. This is consistent with the observation reported in the work of mHuBERT and AudioLM. If we remove the duration model from the U2S, the WER increases dramatically. Our guess is that the unit itself does not contain as much duration information as the phonemes. Therefore the duration model is essential when using unsupervised units.

We further train our own multilingual HuBERT model (mHuBERT_zh_en) with a combination of Chinese and English data. The model size is the same as the HuBERT-large model in (Hsu et al., 2021). We have observed a substantial reduction in the WER metric when utilizing the semantic units generated from mHuBERT_zh_en. Thus, we believe that a multilingual universal representation model trained with more parameters and data can generate better semantic units. We do not use mHuBERT_zh_en in our S2ST experiment because we need the mHuBERT (Lee et al., 2022c) to run the English→Spanish experiment. The benefit of using mHuBERT_zh_en to the overall S2ST is left for future work.

6 Conclusion and Future Work

This paper presents a new framework for speech-to-speech translation (S2ST) based on semantic units. The framework consists of three LMs: a translation LM, a duration LM and a speech synthesis LM. Through comprehensive experimentation, we provide evidence that our unit-based S2ST system surpasses existing systems in terms of ASR-BLEU, ASV, and naturalness metrics. Importantly, our system demonstrates its effectiveness in scenarios involving unwritten languages, where there is a lack of Spanish text transcripts for reference.

Given the significant impact of semantic unit quality on our system's performance, future research will focus on improving the generation of a more refined set of discrete units. We aim to explore techniques and methodologies that can contribute to enhancing the quality and diversity of the generated semantic units. Additionally, we plan to investigate how the system's performance can be further enhanced by scaling up parameters and expanding the available training data. By exploring the effects of increased model capacity and larger datasets, we anticipate uncovering potential improvements in system accuracy and overall effectiveness.

REFERENCES

Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, et al. Findings of the iwslt 2023 evaluation campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pp. 1–61, 2023.

Jason Baldridge. *Verbmobil: Foundations of Speech-to-Speech Translation*, by wolfgang wahlster (editor). springer, 2000. ISBN 3-540-67783-6. price £44.50 (hardback). xii+679 pages. *Nat. Lang. Eng.*, 10(2):200–204, 2004. doi: 10.1017/S1351324904233435. URL https://doi.org/10.1017/S1351324904233435.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2023. doi: 10.1109/TASLP.2023.3288409.

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.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. Gigaspeech: An evolving, multi-domain ASR corpus with 10, 000 hours of transcribed audio. In Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlícek (eds.), *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pp. 3670–3674. ISCA, 2021. doi: 10.21437/Interspeech.2021-1965. URL https://doi.org/10.21437/Interspeech.2021-1965.

Yutian Chen, Yannis Assael, Brendan Shillingford, David Budden, Scott Reed, Heiga Zen, Quan Wang, Luis C. Cobo, Andrew Trask, Ben Laurie, Caglar Gulcehre, Aäron van den Oord, Oriol Vinyals, and Nando de Freitas. Sample efficient adaptive text-to-speech. In *ICLR*, 2019.

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)*, pp. 244–250. IEEE, 2021.

Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, and Junichi Yamagishi. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6184–6188. IEEE, 2020.

Qianqian Dong, Fengpeng Yue, Tom Ko, Mingxuan Wang, Qibing Bai, and Yu Zhang. Leveraging pseudo-labeled data to improve direct speech-to-speech translation. In *Interspeech 2022*, 2022.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. Prompting large language models with speech recognition abilities. *arXiv preprint arXiv:2307.11795*, 2023.

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.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31, 2018.

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. In Gernot Kubin and Zdravko Kacic (eds.), *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pp. 1123–1127. ISCA, 2019. doi: 10.21437/Interspeech.2019-1951. URL https://doi.org/10.21437/Interspeech.2019-1951.

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*, pp. 10120–10134. PMLR, 2022a.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. Cvss corpus and massively multilingual speech-to-speech translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 6691–6703, 2022b.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, Tatiana Likhomanenko, Gabriel Synnaeve, Armand Joulin, Abdelrahman Mohamed, and Emmanuel Dupoux. Libri-light: A benchmark for ASR with limited or no supervision. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pp. 7669–7673. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9052942. URL https://doi.org/10.1109/ICASSP40776.2020.9052942.

Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. Transformer-based direct speech-to-speech translation with transcoder. In *IEEE Spoken Language Technology Workshop, SLT 2021, Shenzhen, China, January 19-22, 2021*, pp. 958–965. IEEE, 2021. doi: 10.1109/SLT48900.2021.9383496. URL https://doi.org/10.1109/SLT48900.2021.9383496.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu-Anh Nguyen, Morgane Rivière, Abdelrahman Mohamed, Emmanuel Dupoux, et al. Text-free prosody-aware generative spoken language modeling. *arXiv preprint arXiv:2109.03264*, 2021.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021.

Alon Lavie, Alex Waibel, Lori S. Levin, Michael Finke, Donna Gates, Marsal Gavaldà, Torsten Zeppenfeld, and Puming Zhan. Janus-iii: speech-to-speech translation in multiple languages. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '97, Munich, Germany, April 21-24, 1997*, pp. 99–102. IEEE Computer Society, 1997. doi: 10.1109/ICASSP.1997.599557. URL https://doi.org/10.1109/ICASSP.1997.599557.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu. Direct speech-to-speech translation with discrete units. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 3327–3339. Association for Computational Linguistics, 2022a. doi: 10.18653/v1/2022.acl-long.235. URL https://doi.org/10.18653/v1/2022.acl-long.235.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, et al. Direct speech-to-speech translation with discrete units. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3327–3339, 2022b.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, et al. Textless speech-to-speech translation on real data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 860–872, 2022c.

Zhaoyu Liu and Brian Mak. Cross-lingual multi-speaker text-to-speech synthesis for voice cloning without using parallel corpus for unseen speakers. *arXiv preprint arXiv:1911.11601*, 2019.

Chutong Meng, Junyi Ao, Tom Ko, Mingxuan Wang, and Haizhou Li. Cobert: Self-supervised speech representation learning through code representation learning. In *Interspeech 2023*, 2023.

Satoshi Nakamura, Konstantin Markov, Hiromi Nakaiwa, Gen-ichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, Jinsong Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto. The ATR multilingual speech-to-speech translation system. *IEEE Trans. Speech Audio Process.*, 14(2):365–376, 2006. doi: 10.1109/TSA.2005.860774. URL https://doi.org/10.1109/TSA.2005.860774.

Tu Anh Nguyen, Benoît Sagot, and Emmanuel Dupoux. Are discrete units necessary for spoken language modeling? *IEEE J. Sel. Top. Signal Process.*, 16(6):1415–1423, 2022. doi: 10.1109/JSTSP.2022.3200909. URL https://doi.org/10.1109/JSTSP.2022.3200909.

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

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)*, pp. 5206–5210. IEEE, 2015.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Kun Song, Yi Ren, Yi Lei, Chunfeng Wang, Kun Wei, Lei Xie, Xiang Yin, and Zejun Ma. Styles2st: Zero-shot style transfer for direct speech-to-speech translation. *arXiv preprint arXiv:2305.17732*, 2023.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Speech-to-speech translation between untranscribed unknown languages. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*, pp. 593–600. IEEE, 2019. doi: 10.1109/ASRU46091.2019.9003853. URL https://doi.org/10.1109/ASRU46091.2019.9003853.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Mirjam Wester and Hui Liang. The emime mandarin bilingual database. Technical report, The University of Edinburgh, 2011.

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 *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pp. 6182–6186. IEEE, 2022. doi: 10.1109/ICASSP43922.2022.9746682. URL https://doi.org/10.1109/ICASSP43922.2022.9746682.

Chen Zhang, Xu Tan, Yi Ren, Tao Qin, Kejun Zhang, and Tie-Yan Liu. Uwspeech: Speech to speech translation for unwritten languages. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 14319–14327. AAAI Press, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17684.

Dong Zhang, Rong Ye, Tom Ko, Wang Mingxuan, and Zhou Yaqian. Dub: Discrete unit back-translation for speech translation. In *Findings in ACL 2023*, 2023a.

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

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

A APPENDIX

A.1 DATASETS

We use a set of several datasets to train U-XLM model: GigaS2S, WenetS2S, LibriLight and some in-house ASR, MT datasets. Table 7 shows the detailed descriptions and statistics.

Type	Dataset	Language	Size	Domain
ASR	LibriLight	En	60K hours	audiobook
	In-house	Zh	60K hours	-
MT	In-house	Zh $\leftrightarrow$ En	44M sents	-
S2S	GigaS2S	En $\rightarrow$ Zh	10K hours	audiobook, podcasts, youtube
	WenetS2S	Zh $\rightarrow$ En	10K hours	youtube

Table 7: Training data of U-XLM model. "-" means the dataset doesn't belong to some specific domains.

A.2 MODEL SETTINGS

A.2.1 S2UT

In the S2UT front-end, U-XLM's model architecture is a unidirectional Transformer decoder consisting of 48 layers with hidden size 1600, feed-forward network (FFN) size 6400, and 25 attention heads. The total parameters are 1.6B. U-XLM is trained on 8/32 NVIDIA TESLA A100 80GB GPUs with a batch size of 3072 tokens per GPU for 500k steps.

A.2.2 U2S

In the U2S back-end, the U-SLM consists of 12 transformer layers. Each of these layers comprises 16 attention heads, an attention dimension of 1024, and an FFN dimension of 4096 in both the autoregressive (AR) model and non-autoregressive (NAR) model. We train the models using 8 NVIDIA TESLA A100 80GB GPUs, with a batch size of 8 utterances per GPU for 800k steps. Training for all steps takes about 5 days.