Speech-to-Speech Translation with Discrete-Unit-Based Style Transfer

Yongqi Wang, Jionghao Bai, Rongjie Huang, Ruiqi Li, Zhiqing Hong, Zhou Zhao
Zhejiang University
cyanbox@zju.edu.cn

Abstract

Direct speech-to-speech translation (S2ST) with discrete self-supervised representations has achieved remarkable accuracy, but is unable to preserve the speaker timbre of the source speech. Meanwhile, the scarcity of high-quality speaker-parallel data poses a challenge for learning style transfer during translation. We design an S2ST pipeline with style-transfer capability on the basis of discrete self-supervised speech representations and codec units. The acoustic language model we introduce for style transfer leverages self-supervised in-context learning, acquiring style transfer ability without relying on any speaker-parallel data, thereby overcoming data scarcity. By using extensive training data, our model achieves zero-shot cross-lingual style transfer on previously unseen source languages. Experiments show that our model generates translated speeches with high fidelity and speaker similarity. [1]

1 Introduction

Speech-to-speech translation (S2ST) aims to translate spoken utterances from one language to another, which can bring immense convenience to international communication. Compared to conventional cascaded systems comprising ASR, text translation, and TTS models (Lavie et al., 1997; Nakamura et al., 2006; Wahlster, 2013), direct S2ST models without intermediate text generation have a more concise pipeline with less computation cost and error propagation, and also facilitates application to unwritten languages, and thus spark widespread interest in the community.

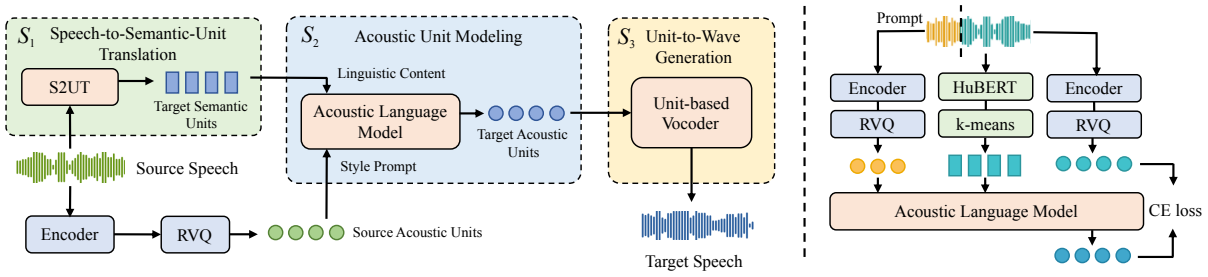
Mainstream approaches of direct S2ST (Lee et al., 2022, 2021; Huang et al.; Popuri et al., 2022) utilize discrete speech representation from self-supervised models (such as HuBERT (Hsu et al., 2021)) as prediction target, and then use them to reconstruct the waveform. Such representation eliminates speaker identity and prosody of the speeches and retains only semantic contents, which simplifies the target distribution and makes the translation less challenging. However, it also has the drawback of losing the style information of the source speech. Extra voice conversion systems are needed if users want to keep the source speaker timbre, which may cause degradation in audio quality.

Some works propose direct S2ST with style transfer (Jia et al., 2021; Song et al., 2023). These methods depend on paired data that source and target speech share the same speakers. However, such data from the real world is extremely scarce as it requires a large number of multilingual speakers, while simulated data from multilingual TTS systems suffers from less diversity and extra data collection costs. Recent large-scale S2ST models (Rubenstein et al., 2023; Barrault et al., 2023) have also incorporated the capability of style transfer, yet their sub-modules are highly coupled and are difficult to apply to other S2ST models.

Inspired by recent progress in spoken language models (Borsos et al., 2023; Wang et al., 2023), we propose a novel approach for direct S2ST with the ability of cross-lingual style transfer, and does not rely on any speaker-parallel data. We utilize two types of discrete representations, namely semantic and acoustic units, from a self-supervised speech model and a neural codec, separately. Our method encompasses three stages: 1) speech-to-semantic-unit translation, which translates source speech to target semantic units; 2) acoustic unit modeling, which generates target acoustic units from translated semantic units using style information in the source speech; and 3) unit-to-wave generation, which reconstructs high-fidelity translated speech from the acoustic units. The modules of the three stages are trained independently and decoupled from each other, allowing our framework to

[1] Audio samples are available at http://stylelm.github.io/

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 34–41



(a) Overall inference pipeline

(b) Training of S_2 stage

Figure 1: We propose an S2ST approach with style transfer based on discrete representations from a self-supervised speech model and a neural codec. Figure (a) shows the inference pipeline of our method; figure (b) illustrates the self-supervised training process of the acoustic language model of S_2.

be applied to various existing speech-to-unit translation models.

For the acoustic unit modeling stage, we introduce an acoustic language model. It employs a self-supervised training approach and learns style transfer through in-context learning, which relies on no speaker-parallel data, and thus addresses the issue of data scarcity. By utilizing extensive training data, our model achieves zero-shot cross-lingual style transfer with source languages not included in the training. Experiments show that our model generates results with superior audio quality and style similarity while maintaining accurate content to a good extent.

Our contributions can be summarized as follows:

- We propose an S2ST approach with cross-lingual style transfer capability, even on previously unseen source languages.
- By employing self-supervised training, our model does not rely on any speaker-parallel data, thus addressing the issue of data scarcity.
- The decoupling nature of the sub-modules enables our framework to be adopted by various existing speech-to-unit translation models.
- Experiments show that our method generates translated speeches with high quality and style similarity.

2 Method

The overall inference pipeline of our method is illustrated in Fig.1 (a). Our method comprises three consecutive stages, utilizing two distinct types of discrete units: 1) speech-to-semantic-unit translation stage S_1, which converts source audio into semantic units of the translated speech; 2) acoustic unit modeling stage S_2, generating target acoustic units conditioned on the semantic output from the preceding stage and the acoustic units of the source speech as style prompt; 3) unit-to-wave generation stage S_3, producing translated speech that maintains consistent style with the source. We provide details about these two types of units and the three stages in the following subsections.

2.1 Semantic and Acoustic Units

Discrete HuBERT (Hsu et al., 2021) units obtained from the clustering of self-supervised speech representations are shown (Lee et al., 2021; Huang et al.) to be effective in providing semantic content information and are widely adopted in S2ST as prediction target (Lee et al., 2022, 2021; Huang et al.; Popuri et al., 2022). HuBERT encodes the target speech into continuous representations with a frame length of 20 ms, and these representations are then discretized with the k-means algorithm to get the semantic units.

On the other hand, audio codec models with encoder-decoder architecture such as SoundStream (Zeghidour et al., 2021) have recently shown outstanding performance in learning acoustic information. Such a codec model can produce discrete representations (i.e. the acoustic units) of audio by employing a convolutional encoder followed by a residual vector quantizer. These representations contain detailed acoustic information and can be used to reconstruct waveforms with the corresponding decoder or an additional vocoder.

2.2 Speech-to-Semantic-Unit Translation

The speech-to-semantic-unit translation stage generates translated semantic units conditioned on source speech input, achieving translation of linguistic content. Various models (Lee et al., 2022;

Huang et al.; Popuri et al., 2022) have been proposed for this procedure. These models share a common basic architecture of a convolutional speech encoder followed by an encoder-decoder architecture based on a transformer (Vaswani et al., 2017) or conformer (Gulati et al., 2020). Due to the decoupling nature of the sub-modules of the three stages, we have the flexibility to adopt different S2UT models in this stage, and we attempted two of them in our experiments (See Section 3.1).

2.3 Acoustic Unit Modeling

The acoustic unit modeling stage S_2 generates translated acoustic units from semantic tokens and style prompts. The core component of S_2 is an acoustic language model, which is basically a decoder-only transformer. Specifically, we adopt UniAudio (Yang et al., 2023) as the acoustic language model, which is proven to be an effective autoregressive audio generation model. Details of the model architecture are provided in Appendix B.1. The model takes a prefix sequence formed by concatenating acoustic unit sequence $\mathbf{a_p}$, which serves as a style prompt, and the target semantic sequence $\mathbf{s}$, and generates the target acoustic sequence $\mathbf{a}$ with autoregressive sampling. This procedure can be formulated as

$$p\left(\mathbf{a} \mid \mathbf{a_p}, \mathbf{s}; \theta_{AR}\right) = \prod_{t=1}^{T} \prod_{c=1}^{C} p\left(\mathbf{a}_t^c \mid \mathbf{a}_{<t}, \mathbf{a}_t^{<c}, \mathbf{a_p}, \mathbf{s}; \theta_{AR}\right) \tag{1}$$

The entire sequence is in the format of $[\mathbf{a_p}|\mathbf{s}|\mathbf{a}]$, with a separator token between each pair of adjacent parts. 3 codebooks are used for $\mathbf{a_p}$ and $\mathbf{a}$.

The training procedure of S_2 is illustrated in Figure 1(b). It adopts a self-supervised training paradigm, where the first three seconds of each audio sample is truncated as prompt, and the acoustic language model is trained to predict the acoustic units of the remaining part conditioned on its semantic units and the prompt acoustic units with cross-entropy loss. This in-context learning approach enables the model to grasp the correspondence in acoustic characteristics between the two parts and acquire style transfer ability. During inference, we use semantic tokens from the previous stage and acoustic units of source speech as the style prompt to realize cross-lingual style transfer.

2.4 Unit-to-Wave Generation

In the waveform generation stage S_3, we adopt a GAN-based unit vocoder to map the target acoustic units to high-fidelity waveforms. Our vocoder is derived from BigVGAN (Lee et al.), with a generator built from a set of look-up tables (LUT) that embed the discrete units, and a series of blocks composed of transposed convolution and a residual block with dilated layers. Multi-period discriminator (MPD) and multi-resolution discriminator (MRD) are used for adversarial training.

3 Experiments

3.1 Setup

Datesets We use two language pairs in the CVSS dataset (Jia et al., 2022) as the translation benchmark, which are French-English (Fr-En) and Spanish-English (Es-En). For S_2 and S_3 stages, we use the *unlab-60k* subset of Libri-Light (Kahn et al., 2020) to train the acoustic language model, and use LibriTTS (Zen et al., 2019) to train the SoundStream model and the vocoder. All audio is processed at a 16 kHz sampling rate. We provide more details about the datasets in Appendix A.

Model Configurations We apply the publicly available multilingual HuBERT (mHuBERT) model[2] with the k-means model of 1000 clusters for the 11th-layer features[3] and train a SoundStream model with a size of 1024 for each codebook and an overall downsampling rate of 320. For stage S_1, we train an S2UT-conformer for Fr-En following (Lee et al., 2022), and follow the model in Popuri et al. (2022) for Es-En but without mbart-decoder initialization. The decoder-only transformer of S_2 has about 760M parameters, with details of its configurations provided in Appendix B.2.

Baselines Considering that previous S2ST models with style transfer (Jia et al., 2021; Song et al., 2023; Rubenstein et al., 2023; Barrault et al., 2023) either differ from ours in settings or are not open-sourced, we mainly compare our model with S2UT models used in S_1 followed by a single-speaker vocoder[4], and cascaded pipelines formed by appending various voice conversion models after the vocoder, which are PPG-VC (Liu et al.,

[2] https://dl.fbaipublicfiles.com/hubert/mhubert_base_vp_en_es_fr_it3.pt

[3] https://dl.fbaipublicfiles.com/hubert/mhubert_base_vp_en_es_fr_it3_L11_km1000.bin

[4] https://github.com/facebookresearch/fairseq/blob/d9a627082fd03ec72a27a31a4e56289bfcb2e4e4/examples/speech_to_speech/docs/textless_s2st_real_data.md#unit-based-hifi-gan-vocoder, English version

ID	Model	BLEU (Fr-En) (↑)	BLEU (Es-En) (↑)	SIM (↑)	MOS (↑)	SMOS(↑)
1	S2UT	18.08	23.78	/	3.73 ± 0.05	/
2	S2UT + PPG-VC	17.03	23.03	0.69	3.37 ± 0.07	3.30 ± 0.06
3	S2UT + NANSY	18.21	23.48	0.68	3.56 ± 0.06	3.47 ± 0.05
4	S2UT + YourTTS	16.23	21.09	0.69	3.74 ± 0.05	3.60 ± 0.06
5	Ours	16.30	22.00	**0.73**	3.86 ± 0.06	$\mathbf{3.69 \pm 0.05}$
6	Target Audio (CVSS-C)	84.36	86.48	/	3.92 ± 0.05	/
7	Target Audio (CVSS-T)	80.99	82.12	0.69	3.95 ± 0.05	3.56 ± 0.06

Table 1: Results on translation quality and audio similarity on CVSS dataset.

ID	Model	SIM (↑)	MOS (↑)	SMOS (↑)
1	LibriTTS	0.67	3.84 ± 0.05	3.55 ± 0.05
2	Libri-Light unlab-60k	0.73	3.86 ± 0.05	3.69 ± 0.05
3	+ CVSS source	0.78	3.85 ± 0.05	3.74 ± 0.06

Table 2: Ablation results on different compositions of training data.

2021), NANSY (Choi et al., 2021) and YourTTS (Casanova et al., 2022).

Evaluation Metrics We employ both objective and subjective metrics to measure the model performance in terms of translation accuracy, speech quality, and style similarity with the source speech. For objective evaluation, we calculate the BLEU score between the ASR-transcripts of the translated speech and reference text as well as speaker cosine similarity (SIM). For subjective metrics, we use crowd-sourced human evaluation with 1-5 Likert scales and report mean opinion scores on speech quality (MOS) and style similarity (SMOS) with 95% confidence intervals (CI). More details are provided in Appendix C.

3.2 Results and Analysis

Table 1 summarizes the main experiment results. In terms of audio quality, our model achieves a high MOS of 3.86, surpassing baselines 2-4. This demonstrates the significant advantage of our model in speech naturalness compared to cascaded pipelines with voice conversion models. Moreover, our model gets higher MOS than direct S2UT, indicating that incorporating acoustic unit modeling helps improve the long-term naturalness of speech. On the other hand, our model achieves the highest speaker similarity, with SMOS being 3.69 and SIM being 0.73, which surpasses all three cascaded systems and even the CVSS-T target, demonstrating the outstanding performance in zero-shot cross-lingual style transfer of our model. This can be attributed to the large model size and extensive training data, through which our model acquires strong zero-shot style transfer capability and can generalize effectively to unseen source languages.

In terms of translation accuracy, generally, there is a comprehensive decrease in BLEU scores for 2-5 compared to 1, indicating that additional style transfer processes lead to a loss in semantic content. Compared to PPG-VC and NANSY, YourTTS and our model suffer from lower BLEU scores. We observe that this is due to the acoustic environment transfer capabilities of YourTTS and our S2 stage model, which transfer some of the strong background noise from the source speech into the generated speech, posing a challenge for ASR. Nevertheless, our model still maintains good translation accuracy, with BLEU declination restricted to 1.78 for both Fr-En and Es-En, outperforming the cascaded baseline with YourTTS.

3.3 Ablation Studies

We further conduct ablations on different training data compositions of S_2, and the results are summarized in Table 2. We observe that when using LibriTTS with a smaller size and fewer speakers, there is a significant decrease in SMOS and SIM of 0.14 and 0.06, with only a minor decrease in MOS of 0.02. This suggests that the model's style transfer performance relies on a large amount of speech data from multiple speakers while achieving high-quality speech generation does not require as much data.

We also add part of the speech from the CVSS source to the training data to examine the model performance on unseen / seen speakers. We observe a gap of 0.05 for both SIM and SMOS. This indicates that our model's zero-shot style similarity still lags behind that of seen speakers. This gap can be narrowed by using a training corpus with more speakers.

4 Conclusions

We propose an S2ST approach with style transfer capability by adopting an acoustic language model that learns style transfer through in-context learning. By adopting self-supervised training and large-scale training data, our method addresses the scarcity of speaker-parallel data and achieves cross-lingual style transfer with unseen source languages. Experiments indicate that our approach achieves outstanding results in terms of speech quality and style similarity while keeping good translation accuracy.

5 Limitations and Potential Risks

Despite that our model excels in style transfer and generating high-quality translated speech, it still suffers from several limitations: 1) Our evaluation (especially the objective evaluation) of style transfer capability mainly focuses on the global speaker timbre, and we have not yet delved deeply into other stylistic characteristics such as prosody and emotion. We leave the exploration of these aspects for future work. 2) The large model size and the autoregressive generation paradigm may lead to efficiency issues, such as long inference latency. 3) The BLEU scores heavily depend on the ASR quality, which may not accurately reflect the speech translation performance. Future directions could be improving ASR quality or exploring other evaluation metrics without reliance on ASR models. Besides, due to the speaker timbre transfer capability of our model, it may be misused to disinform, defame, or commit fraud. We will add some constraints to guarantee people who use our code or pre-trained model will not use the model in illegal cases.

Acknowledgements

This work is supported by National Key R&D Program of China under Grant No.2022ZD0162000, National Natural Science Foundation of China under Grant No. 62222211 and No.62072397.

References

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

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

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

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

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

Rongjie Huang, Jinglin Liu, Huadai Liu, Yi Ren, Lichao Zhang, Jinzheng He, and Zhou Zhao. Transpeech: Speech-to-speech translation with bilateral perturbation. In *The Eleventh International Conference on Learning Representations*.

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

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

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

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

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, et al. 2022. 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)*, pages 3327–3339.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, et al. 2021. Textless speech-to-speech translation on real data. *arXiv preprint arXiv:2112.08352*.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. In *The Eleventh International Conference on Learning Representations*.

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

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

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Sravya Popuri, Peng-Jen Chen, Changhan Wang, Juan Pino, Yossi Adi, Jiatao Gu, Wei-Ning Hsu, and Ann Lee. 2022. Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation. *arXiv preprint arXiv:2204.02967*.

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

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

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

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

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*.

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

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

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

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

A Datasets

In this section, we provide details of the translation benchmark dataset and the corpora for training S_2 and S_3 models.

CVSS CVSS (Jia et al., 2022) is an S2ST benchmark dataset derived from the CoVoST 2 (Wang et al., 2020) speech-to-text translation corpus by synthesizing the translation text into speech using TTS systems. It comprises two sub-versions of CVSS-C and CVSS-T, where the target speech in CVSS-C is generated by a single-speaker TTS system while that of CVSS-T is generated by a multi-speaker TTS system with speaker timbre transferred from the source speech. We use CVSS-C for training and evaluating the translation models, and provide results of ground truth target audios in CVSS-T as a reference for style transfer performance.

Libri-Light Libri-Light is a large-scale corpus containing unlabelled speech from audiobooks in English. The *unlab-60k* subset we use consists of 57.7k hours of audio with 7,439 speakers.

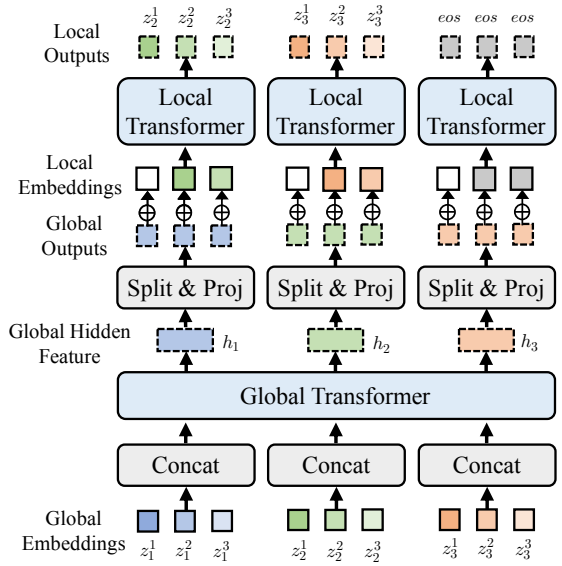


Figure 2: The multi-scale architecture of UniAudio used for the S_2 stage model.

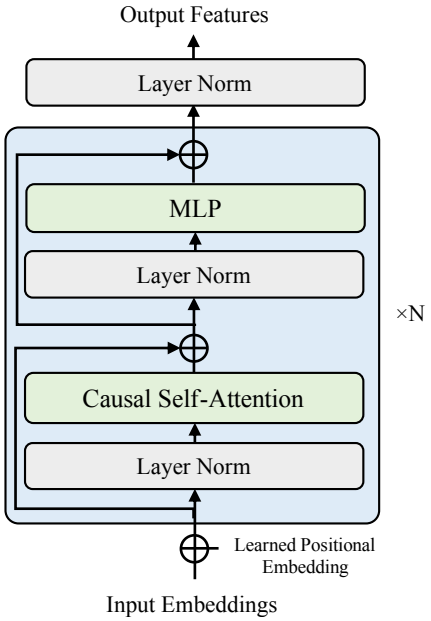


Figure 3: Structure of the global transformer.

LibriTTS LibriTTS is a multi-speaker English TTS dataset. It comprises 585.5 hours of audio with 2,456 speakers.

B Model Settings

B.1 S_2 Model Architecture

UniAudio (Yang et al., 2023) is a decoder-only transformer with an end-to-end differentiable multi-scale architecture to facilitate the modeling of long sequences. It has a hierarchical structure consisting of a global transformer and a local one. Figure 2 illustrates its multi-scale design. This model has exhibited remarkable capabilities in audio synthesis and modeling intrinsic relationships between acoustic and other modalities, as well as high efficiency in generating long sequences based on sub-quadratic self-attention. In this work, we adopt UniAudio as our S_2 stage model.

The architecture of the global transformer is illustrated in Figure 3. The local transformer shares the same structure as the global one with two differences: 1) the local transformer has no positional embedding, and 2) there is a linear lm-head appended to the top for token prediction.

B.2 Model Parameters

We provide hyperparameters of our S_2 and S_3 stage models in Table 3. We also refer the readers to the original papers (Lee et al., 2022; Popuri et al., 2022) for details of S_1 models used. Each sub-module is trained with 4 NVIDIA-V100 GPUs for about a week.

C Evaluation Metrics

For translation accuracy, we use an open-sourced ASR model in *fairseq* [5] (Ott et al., 2019) to transcribe the audios and then calculate the BLEU score between the transcripts and the reference text. For speaker similarity, we use Resemblyzer[6], which is a public-available speaker encoder to extract speaker embeddings of the synthesized and source speech and calculate their cosine similarity.

Our subjective evaluation tests are crowd-sourced and conducted via Amazon Mechanical Turk. For audio quality evaluation, we ask the testers to examine the audio quality and naturalness. For style similarity, we instruct the testers to evaluate the style similarity between the synthesized and source speech while ignoring the content. The testers rate scores on 1-5 Likert scales. We provide screenshots of the testing interfaces in Figure 4 and 5. Each data item is rated by 2 testers, and the testers are paid $8 hourly.

Due to the large cost of conducting voice conversion and evaluation on the whole test split, we randomly sample 488 items from each language pair for evaluation, which represents approximately

[5] https://github.com/facebookresearch/fairseq/tree/main/examples/speech_to_speech/asr_bleu

[6] https://github.com/resemble-ai/Resemblyzer

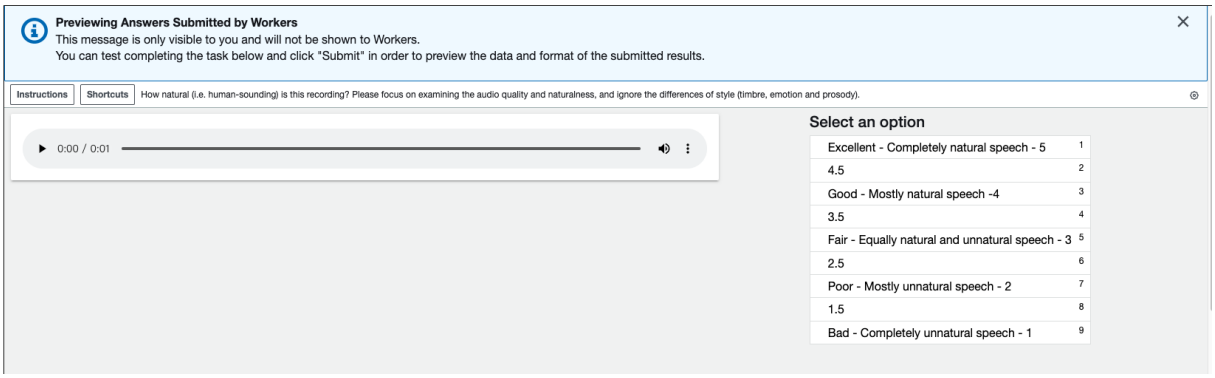


Figure 4: Screenshot of MOS testing.

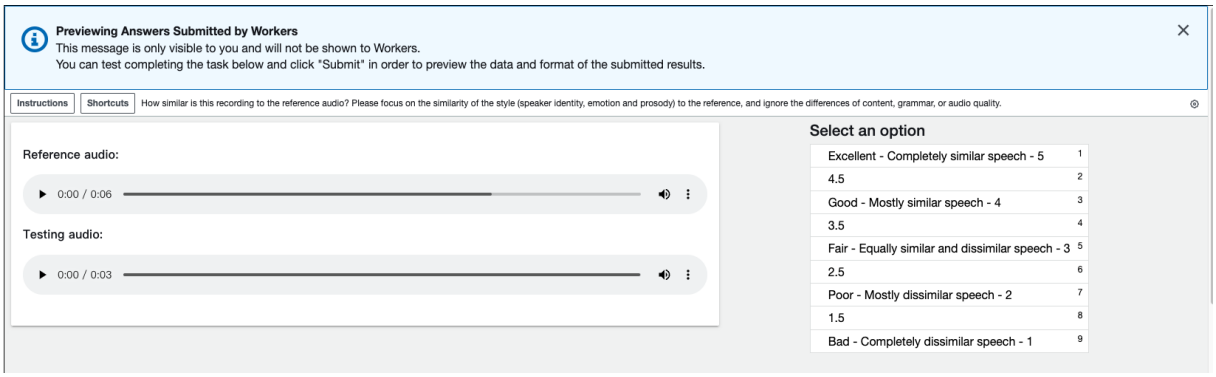


Figure 5: Screenshot of SMOS testing.

3% of the test set.

Model	Hyperparameter	
Acoustic Language Model	Global Layers	20
	Local Layers	6
	Hidden Dim	1,536
	Attention Headers	16
	FFN Dim	6,144
	Number of Parameters	763.1M
Unit Vocoder	Upsample Rates	[5,4,2,2,2,2]
	Hop Size	320
	Upsample Kernel Sizes	[9,8,4,4,4,4]
	Number of Parameters	121.6M

Table 3: Hyperparameters of S_2 and S_3 Stage Models.