Multi-Task Self-Supervised Learning Based Tibetan-Chinese Speech-to-speech Translation

Rouhe Liu
Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE
Minzu University of China
School of Information Engineering
Minzu University of China
Beijing, China
21302003@muc.edu.cn

Yue Zhao (Corresponding author)
Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE
Minzu University of China
School of Information Engineering
Minzu University of China
Beijing, China
zhaoyueso@muc.edu.cn

Xiaona Xu
Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE
Minzu University of China
School of Information Engineering
Minzu University of China
Beijing, China
xuxiaona@muc.edu.cn

Abstract*—Speech-to-speech translation tasks are commonly tackled by using a three-level cascade system which comprises of speech recognition, machine translation, and speech synthesis. However, this approach suffers from the drawback of error accumulation at each stage. In contrast, the direct speech-to-speech translation model directly converts speech from the source language to the target language without relying on intermediate text generation, thereby avoiding the issue of incorrect transmission in cascading systems. Currently, there exist two categories for direct speech-to-speech translation methods. The first involves mapping the Mel-spectrogram of the source language speech to the Mel-spectrogram of the target language speech. However, this method often encounters challenges in convergence and producing the audible speech for the target language. The second type of methods is to learn a self-supervised discrete representation of the target language using an unlabeled speech corpus. This method entails training a sequence-to-sequence model on a real-world dataset, which then maps the source language speech to the discrete representation of the target language. Finally, a separately trained vocoder is utilized to convert the discrete unit sequence into a speech waveform. Given the limited availability of large-scale Tibetan-Chinese parallel speech corpora, this work adopts the second method to model Tibetan-Chinese speech-to-speech translation tasks. Additionally, a multi-task learning framework is designed in this work to enhance the performance of the speech translation model. Experimental results demonstrate that the Tibetan-Chinese speech-to-speech translation model based on multi-task self-supervised learning outperforms both the model based on spectrogram mapping and the single-task self-supervised learning model in terms of achieving a higher BLUE value.

***Keywords*—Tibetan-Chinese speech-to-speech translation, Self-supervised learning, Multitask learning, Transformer**

I. INTRODUCTION

Generally, speech-to-speech translation can be achieved through a common three-step process including automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS) synthesis [1-2]. In recent years, speech-to-text translation (S2TT) [3] or text-to-speech translation (T2ST) [4] has simplified the speech-to-speech translation pipeline into two stages. Although these speech-to-speech translation systems have gained a solid foundation after years of research and development and are gradually moving towards commercial applications, they aim to eliminate communication barriers between people speaking different languages and provide access to multimedia content in different languages. However, they still have three disadvantages. Firstly, any step in the cascade speech-to-speech translation cannot achieve 100% accuracy, leading to error propagation. Errors generated by ASR can affect the quality of MT and TTS. Secondly, due to the multiple translation processes involved in cascading speech translation, the translation time is longer, and reducing the time cost of any step does not significantly improve the overall time loss of the system. Thirdly, each process requires support from one to two models, resulting in high storage overhead and a heavy system load.

Recently, many scholars have begun to explore the establishment of a speech-to-speech Translation (S2ST) model, which aims to directly translate source language speech into target language speech without the need for text generation as an intermediate step. Therefore, it can effectively reduce the three major drawbacks in cascade S2ST systems, which has important research significance. One method for modeling S2ST is to train a model that directly maps the Mel spectrum of one language to the Mel spectrum of another language. However, this method requires the model not only to learn how to align two languages (such as MT) but also to learn the appropriate acoustic and linguistic features of both languages (such as ASR and TTS). The training of model requires the large scale of parallel corpora.

In the year of 2021, the works of [21-23] achieved success in speech self-supervised learning. These works show that discretization speech units obtained from clustering of self-supervised speech representation can achieve higher quality speech synthesis and speech generation, better speech data modeling, and learn from audio to generate natural and smooth spoken language expression.

Therefore, our work adopts the speech self-supervised learning method, and explore to solve the problem of target speech continuous modeling in direct Tibetan-Chinese S2ST by predicting the self-supervised discrete representation of target speech of Chinese. At the same time, due to the lack of Tibetan-Chinese parallel corpus, we designed a multi-task learning framework, which helps improve the performance of the main task of Tibetan-Chinese speech translation through auxiliary Tibetan speech recognition tasks. We conducted the S2ST model on a real parallel dataset of Tibetan-Chinese speech, and the experimental results showed the effectiveness of our method.

979-8-3503-3078-6/23/$31.00

II. RELATED WORK

The existing works of modelling speech-to-speech translation have been conducted on the following three aspects.

A. Traditional Speech-to-Speech Translation

The traditional S2ST system usually follows a three-level cascade approach, namely ASR+MT+TTS. Over time, it evolved into a two-stage cascade system, combining end-to-end S2TT and TTS [1-2], and another approach involves the combination of ASR and end-to-end T2ST.

Most research on two-stage cascade has focused on end-to-end S2TT, which shows promise in addressing the drawbacks of error accumulation in ASR+MT integrated systems. By utilizing multitask learning or pre-training model, researchers have been able to overcome the challenges of data scarcity and achieve better performance than ASR and MT integrated models.

In addition, the study of T2ST emphasizes the accurate transmission of source language information, such as prosodic and word-level stress, which plays a crucial role in enhancing the naturalness, fluency, and comprehensibility of S2ST systems.

B. Direct Speech-to-Speech Translation

Michelle Guo et al. [7] have explored direct S2ST tasks using a 70-word corpus for translating English speech to Chinese speech and speaker recognition. Although the experimental results showed promising translation outcomes, it was observed that the limited number of words in the corpus and the big overlap between the test set and the training set may have influenced the results. Nonetheless, this work demonstrated the potential of sequence-to-sequence models in direct S2ST tasks.

Translatotron [5] is an attention-based sequence-to-sequence framework that directly predicts the Mel spectrum as the model output, mapping the Mel spectrum of the source speech to the Mel spectrum of the target speech. However, this model still lags behind the S2TT+TTS cascade system in performance. Moreover, Tjandra et al. and Jia et al. [5,18] noted that the model struggles to fully converge and generate audible target speech without the use of auxiliary tasks.

Kano et al. [17] introduced a Transformer-based version of Translatotron, where ASR, MT, and TTS models were pre-trained separately. The pre-trained ASR encoder was employed for encoding the source language speech, and ASR, MT, and TTS decoders were used for generating source language text, target language text, and target language speech, respectively. Subsequently, they employed a transcoder to connect the ASR encoder with the MT decoder, and another transcoder to link the MT decoder with the TTS decoder. However, during the inference process, both the ASR and MT decoders are required to complete the entire sequence decoding, leading to the loss of advantages of a direct S2ST system.

Tjandra et al. [18] and Zhang et al. [3] employed a vector quantised variational auto-encoder (VQ-VAE) [15] to convert target speech into discrete representations and trained sequence-to-sequence models to translate speech into discrete units, as well as inverters to transform units back into speech, thereby constructing a direct speech-to-speech translation system for unwritten languages. These systems either did not use any text data [18] or employed text data from other languages [3] for training. Recently, Lee et al. [6] utilized the HuBERT self-supervised speech model [10] and applied the knowledge from speech-to-text modeling to speech-to-speech translation. They investigated encoding target speech into self-supervised discrete representations to train a direct speech-to-speech translation model and designed a multi-task learning framework that combines speech and text training, enabling the model to generate both speech and text outputs simultaneously. However, the parallel speech data used for training this model consists of source-language speech from a single speaker and synthesized target-language speech..

C. Data Scarcity Issues

A common challenge faced during the training of S2ST models is data scarcity. To tackle this issue, Jia et al. [11] employed high-quality English TTS to create an X→En S2ST dataset comprising 21 synthesized target speeches. In order to create a S2ST dataset with real speech, Wang et al. [12] aligned ASR transcripts for more than 100 languages to create a S2ST dataset with real speech. Duquenne et al. [13] proposed an automatic data mining method to perform speech-to-speech mining, generating a S2ST mining dataset between 17 European languages. Additionally, Dong et al. [14] effectively improved the performance of S2ST models by using large-scale pseudo localization tag data. Lee et al. [6] and Kano et al. [17] addressed data scarcity and model convergence issues by incorporating multitask learning and pre-training methods.

III. OUR METHOD

In this work, based on the Translatotron [5], we trained a speech-to-unit (S2U) model through multi-task self-supervised learning for Tibetan-Chinese direct S2ST. We adopt the self-supervised representations of HuBERT [10] to generate Chinese speech discrete units for our task. Compared to other unsupervised representations (including VQ-VAE-based representations used in the works of [3,18]), Hubert's self-supervised representation has shown superior performance in ASR [21], oral modeling [22], and TTS [23]. However, due to the lack of parallel S2ST training data, previous work on direct S2ST mostly utilized TTS technology to generate target speech for model training [3,5,17]. However, our work used real S2ST datasets and did not use synthetic audio.

The architecture of our proposed model is shown in Figure 1, which is a Transformer-based sequence-to-sequence model that can be decomposed into three parts. The first is a S2U model with a speech encoder and a discrete unit decoder. The second part is a speech recognition auxiliary task added during the training peocess to promote model learning, which is similar to Translatotron. The third is a separately trained vocoder which is used to convert discrete units into speech waveforms.

A. Speech-to-Unit (S2U) Model

The HuBERT model trained on an unlabeled speech corpus with Chinese as the target language can encode the input speech waveforms into K-means clustering index sequences every 20-milliseconds frame. Learning representations for unannotated speech involves generating K cluster centroids using the K-means algorithm, and representing the target speech by clustering indices for every 20-millisecond segment. Finally, the Chinese target discourse

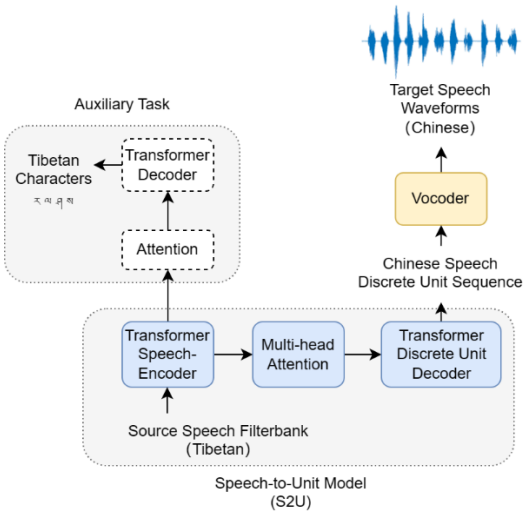

Fig. 1. Model architecture.

y is represented as $[z_1, z_2, \ldots, z_T]$, $z\_i \in \{0,1,\ldots,K-1\}$, $\forall 1 \leq i \leq T$, where T is the number of frames.

The S2U model is constructed on the Transformer proposed by Vaswani et al. [9]. This model comprises a speech encoder and a unit decoder. The speech encoder is built by concatenating a speech downsampling module with a stack of transformer blocks. The downsampling module consists of two one-dimensional convolutional layers, each with stride 2, followed by a gated linear unit activation function. Details of the encoder module are shown in Figure 2. The unit decoder is a stack of transformer blocks. Given the discrete nature of the target sequence, we train the S2U model using cross-entropy loss with labeled smoothing, following the "reduced" strategy from [6], where consecutive sequence of the same unit were collapsed into one single unit, resulting in a series of unique discrete units. This strategy accelerates training and inference and enhances model performance.

B. *Multi-task Learning*

Following the design for unwritten language scenarios from [6], we introduce a speech recognition auxiliary task to help the model converge. This task is applied on intermediate layers of the speech encoder, involving attention mechanisms and transformer decoder modules. The target output of the auxiliary task can be either phonemes, characters, subword units, or any discrete representations of source or target utterances. According to the results in [6], where using characters as targets for the auxiliary task gives 7 BLEU gain compared to phonemes, , we use Tibetan characters as targets for the auxiliary task (as shown in Figure 3) in our work. The auxiliary task is only used during training and not in inference.

C. *Unit-based Vocoder*

For unit-to-speech conversion, we adopt an improved version of the discrete-unit-based HiFi-GAN vocoder [19] proposed by Polyak et al. [23]. For the discrete unit output, we enhance the vocoder with a duration prediction module from Ren et al. [20] to recover the duration for the reduced unit sequence, as shown in Figure 4. This duration prediction module consists of two one-dimensional convolutional layers and a linear layer. We train the enhanced vocoder using the mean squared error (MSE) between the module prediction and the ground truth duration of each unit segment in logarithmic domain, in addition to the generator-discriminator loss from HiFi-GAN.

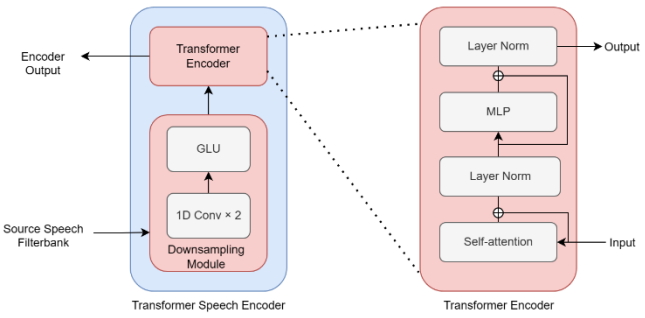

Fig. 2. Speech encoder architecture.

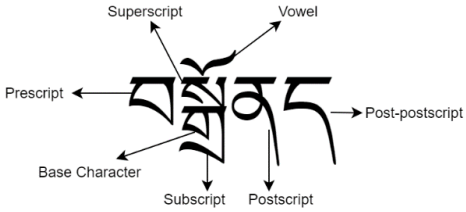

Fig. 3. Tibetan character.

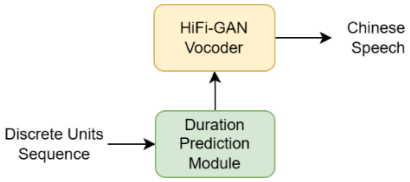

Fig. 4. Unit-based HiFi-GAN vocoder

IV. Experiment

A. *Experimental Data*

Unlike the cascaded S2ST approach, training end-to-end S2ST models requires a large scale of dataset of parallel speech pairs. To evaluate our proposed method, our task utilizes Tibetan-Chinese parallel speech data collected by our research group, as well as English-Chinese parallel speech data from Common Voice [16]. Our Tibetan-Chinese parallel speech corpus covers a wide range of content, including daily conversations and folk culture, spanning a diverse vocabulary that includes common words, proper nouns, and geographical names. Details of the dataset are provided in Table 1.

Due to the compressed MP3 format, the Chinese speech data from Common Voice is converted to the uncompressed WAV format. For standardization, all speech data is downsampled to 16kHz before training subsequent model.

B. *Experimental Setup*

To quantify speech, we adopt an acoustic representation-based approach, learning K-means clustering on acoustic representations. We use the HuBERT-based acoustic model pretrained on Librispeech [8] to discretize speech. Following the ideas from the works [10, 22], we use 100 units from the sixth layer of the HuBERT-Base model to extract discrete units of the target speech.

We apply Tibetan characters or English words as the target output of the auxiliary task, adding a 4-head multi-head attention module and 2 transformer decoder layers on the eighth layer of the encoder, with the same embedding size as the discrete unit decoder. The weight of auxiliary loss is fixed at 8.0.

TABLE I. DETAILS OF PARALLEL SPEECH CORPUS

Tibetan-Chinese parllel speech corpus	***Tibetan***	***Chinese***
Total duration	10h35m	10h5m
Number of sentences	3391	3391
Average duration per sentence	11.24s	10.70s
Audio format	.wav	.wav
Sampling rate	16kHZ	16kHZ
Common Voice(En→Cn)	***English***	***Chinese***
Total duration	20h29m	26h35m
Number of sentences	16825	16825
Average duration per sentence	4.38s	5.68s
Audio format	.wav	.wav
Sampling rate	24kHZ	48kHZ

C. Baseline

For baseline models, we select a LSTM-based Translatotron model and a Transformer-based Translatotron model. To stay close to the original models, we replace the positional sensitive attention mechanism with a 4-head multi-head attention mechanism, with no pretraining or multi-task learning added.

The Transformer-based Translatotron for Tibetan-Chinese speech translation is depicted in Figure 5. Although This model is able to generate rib-like patterns in the spectrogram, as shown in Figure 6, but it encounters challenges in producing the audible speech for target language of Chinese.

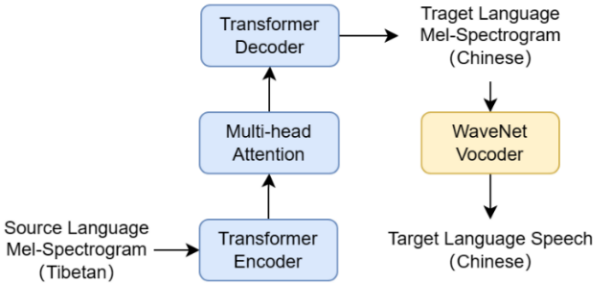

Fig. 5. Transformer-Based Translatotron Model for Tibetan-Chinese S2ST.

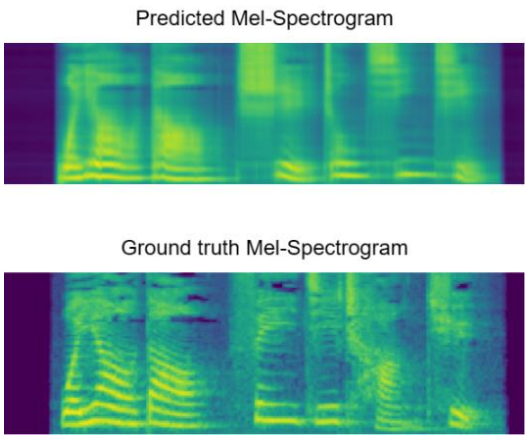

Fig. 6. Predicted Mel-Spectrogram and Ground truth Mel-Spectrogram.

TABLE II. EXPERIMENTAL RESULT

BLEU		***Tibetan-Chinese***		***English-Chinese***	
		dev	test	dev	test
Single-task	LSTM-based Translatotron	0.0	0.0	0.1	0.0
	Transformer-based Translatotron	0.0	0.0	0.16	0.0
	S2U	22.11	21.95	19.39	19.22
Multi-task	S2U	**25.13**	**24.73**	**23.17**	**21.84**

D. Result

The evaluation of experimental results employs the BLEU (bilingual evaluation understudy) score [24], a commonly used metric in machine translation. This score primarily measures the discrepancy between predicted target speech content and reference target speech content, with closer alignment indicating better translation performance.

Table 2 summarizes the results obtained using Tibetan-Chinese parallel speech data and English-Chinese parallel speech data. The corpus is divided into training data, development data, and test data in 6:2:2 ratio. Algorithm tuning is performed on the development data, evaluating the performance of trained models. Finally, the model with the best-performing hyperparameters is evaluated on the test data to determine model performance.

From the experimental results, it is evident that both the LSTM-based Translatotron with single-task and the Transformer-based Translatotron with single-task struggle to generate audible speech for target language. In contrast, the single-task S2U model is capable of direct S2ST for Tibetan-Chinese and English-Chinese. Furthermore, the multi-task S2U model with auxiliary tasks demonstrates effective enhancement in speech translation performance over the single-task S2U model, achieving higher BLEU scores. Multi-task learning can effectively address data sparsity and promote model convergence.

V. CONCLUSION

Direct S2ST serves as a bridge for communication between speakers of different languages, greatly overcoming language barriers and reducing cross-cultural disparities. This research field holds significant research value and developmental potential. In this paper, a direct S2ST model is trained on a real-life Tibetan-Chinese speech dataset. We start by training a discrete quantization autoencoder to generate discrete representations from target Chinese speech features. Next, a sequence-to-sequence model is trained to predict sequences of discrete units from given speech representations. Finally, an independent vocoder was trained to convert discrete unit sequences into target language speech waveforms. This model employs speech recognition as an auxiliary task in conjunction with the speech translation task for multi-task learning, which enhances the performance of Tibetan-Chinese S2ST model. In future work, we will combine pretraining techniques and data augmentation methods to further improve the translation performance of this model.

REFERENCES

[1] 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.

[2] 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.

[3] Alexandre Bérard, Olivier Pietquin, Laurent Besacier, and Christophe Servan. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In NIPS Workshop on end-to-end learning for speech and audio processing.

[4] Chen Zhang, Xu Tan, Yi Ren, Tao Qin, Kejun Zhang, and Tie-Yan Liu. 2020. UWSpeech: Speech to speech translation for unwritten languages. arXiv preprint arXiv:2006.07926.

[5] Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu.2019. Direct speech-to-speech translation with a sequence-to-sequence model. Proc. Interspeech 2019, pages 1123–1127.

[6] Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, XLeveragingutai Ma, Adam Polyak, Yossi Adi, Qing He, Y un Tang, et al 2022a. 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.

[7] Guo M, Haque A, Verma P. End-to-End Spoken Language Translation. arXiv e-prints [Internet]. 2019 April 01, 2019. Available from: https://ui.adsabs.harvard.edu/abs/2019arXiv190410760G.

[8] 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). IEEE, 2015, pp. 5206–5210.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008.

[10] 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. arXiv preprint arXiv:2106.07447.

[11] Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022c. CVSS corpus and massively multilingual speech-to-speech translation. arXiv preprint arXiv:2201.03713.

[12] Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021b. CoVoST2 and massively multilingual speech translation. In Interspeech, pages 2247– 2251.

[13] Paul-Ambroise Duquenne, Hongyu Gong, Benoît Sagot, and Holger Schwenk. 2022b. T-Modules: Translation modules for zero-shot cross-modal machine translation. arXiv preprint arXiv:2205.12216.

[14] Qianqian Dong, Fengpeng Yue, Tom Ko, Mingxuan Wang, Qibing Bai, and Yu Zhang. 2022. Leveraging pseudo-labeled data to improve direct speech-to-speech translation. arXiv preprint arXiv:2205.08993.

[15] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In Proceedings of the 31st International Conference on Neural Information Processing Systems, pages 6309–6318.

[16] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M. and Weber, G. (2020) "Common Voice: A Massively-Multilingual Speech Corpus". Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). pp. 4211—421.

[17] Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura, “Transformer-based direct speech-to-speech translation with transcoder,” in 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021, pp. 958–965.

[18] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura, “Speech-to-speech translation between untranscribed unknown languages,” in 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019, pp. 593–600.

[19] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, “HiFiGAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” Advances in Neural Information Processing Systems, vol. 33, 2020.

[20] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech2: Fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558.

[21] Shu-wen Y ang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al, “SUPERB: Speech processing universal performance benchmark,” arXiv preprint arXiv:2105.01051, 2021.

[22] Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Adelrahman Mohamed, et al, “Generative spoken language modeling from raw audio,” arXiv preprint arXiv:2102.01192, 2021.

[23] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux, “Speech resynthesis from discrete disentangled self-supervised representations,” arXiv preprint arXiv:2104.00355, 2021.

[24] Matt Post, “A call for clarity in reporting BLEU scores,” in Proceedings of the Third Conference on Machine Translation: Research Papers, 2018, pp. 186–191.