Improving Speaker Consistency in Speech-to-Speech Translation Using Speaker Retention Unit-to-Mel Techniques

Rui Zhou* and Akinori Ito† and Takashi Nose‡
* Graduate School of Engineering, Tohoku University, Sendai
E-mail: zhou.rui.p1@dc.tohoku.ac.jp
† Graduate School of Engineering, Tohoku University, Sendai
E-mail: aito.spcom@tohoku.ac.jp
‡ Graduate School of Engineering, Tohoku University,
E-mail: takashi.nose.b7@tohoku.ac.jp

***Abstract*—We propose a Speaker-Consistent Speech-to-Speech Translation (SC-S2ST) system that effectively retains speaker-specific information. While the paradigm of Speech-to-Unit Translation (S2UT) followed by Unit-to-Waveform Vocoder has become a mainstream for End-to-End S2ST systems, due to the substantial semantic content carried by discrete units, this approach primarily captures semantic information and often results in synthesized speech that lacks speaker-specific characteristics such as accent and individual voice qualities. Existing S2UT systems with style transfer face the issue of high inference latency. To address this limitation, we introduced a Speaker-Retention Unit-to-Mel (SR-UTM) framework designed to capture and preserve speaker-specific information. We conducted experiments on the CVSS-C and CVSS-T corpora for Spanish-English and French-English translation tasks. Our approach achieved BLEU scores of 16.10 and 21.68, which are comparable to those of the baseline S2UT system. Furthermore, our SC-S2UT system excelled in preserving speaker similarity. The speaker similarity experiments showed that our method effectively retains speaker-specific information without significantly increasing inference time. These results confirm that our primary approach successfully achieve speaker-consistent speech-to-speech translation.**

I. Introduction

Speech-to-speech translation (S2ST) refers to the process of converting spoken language in one language into spoken language with the same meaning in another language, facilitating communication between speakers of different languages. Traditional cascade S2ST systems operate by first using automatic speech recognition (ASR) to transcribe the source speech, followed by machine translation (MT) to translate the source text into the target text, and finally using text-to-speech (TTS) to synthesize the target speech. More recently, end-to-end speech-to-text (S2T) models[1] have been employed to replace the ASR and MT components, reducing the accumulation of errors.

However, the cascade systems mentioned above face significant challenges, such as long inference times and the large parameter size due to multiple models, making them impractical for industrial applications. Consequently, researchers have shifted their focus to E2E S2ST systems. Translatotron was the first proposed E2E S2ST, utilizing auxiliary recognition and attention-based sequence-to-sequence multitask learning to transform spectrograms from the source language to the target speech spectrograms[2]. Subsequently, Translatotron2, which utilizes a two-pass approach, was proposed[3]. This method establishes attention between the source language acoustic features and the target language phonemes, thereby reducing the complexity of directly predicting the target spectrogram. Additionally, phoneme-level content features are employed during inference, enhancing performance compared to the original Translatotron.

The Translatotron series of S2ST primarily focuses on transforming the spectrogram of the source language into the spectrogram of the target language. However, directly building a translation model between two spectrograms is exceptionally challenging due to the high dimension and complexity of spectrogram data. In recent years, with the advent of self-supervised learning, a new approach has emerged, which involves extracting speech representations using models like HuBERT[4] to capture rich and detailed features of the speech signal. These representations are then clustered to construct discrete units, which serve as a more manageable and semantically meaningful intermediate representation. A combination of unit-to-waveform conversion and advanced vocoders [5] has gradually become mainstream [6], [7].

Some previous works have applied style transfer in direct S2UT[8], [9]. These methods utilize an acoustic language model to generate units containing speaker information, which are then combined with semantic units and finally converted to speech using a vocoder. However, this approach used an auto-regressive method to generate units greatly increases inference time. Additionally, the large parameter size of the Residual Vector Quantization model results in reduced efficiency. To address these issues, we propose our Speaker-Retention Unit-to-Mel framework(SR-U2M), which enhances efficiency.

[0]This work was supported by JST SPRING, Grant Number JPMJSP2114 and JSPS KAKENHI JP 23K20725 and JP24H00085

There are two main motivations for our work. Firstly, we found that mel-spectrograms, compared to units, have higher dimensionality and contain richer information, making it easier to retain speaker information. Secondly, we were inspired by FastSpeech[10], a non-auto-regressive text-to-speech method that quickly synthesizes mel-spectrograms from discrete units.

Therefore, we introduce a novel structure SR-U2M, which integrates the speaker's reference utterance along with the speech units into the model. The SR-U2M framework first extracts speaker information using a speaker adapter, and then embeds this speaker information into the speech units. The combined information is subsequently converted into a mel-spectrogram through a unit-to-mel structure, and finally transformed into speech. This approach not only effectively preserves the speaker's unique characteristics but also significantly enhances inference efficiency.

The rest of this paper is organized as follows. First, we introduce related work in the next section. Then, we describe our SC-S2ST system and SR-U2M in Section III. Following this, we present and analyze our experimental results in Section IV. Section V summarizes with the experimental results and analysis. Finally, Section VI is the conclusion.

II. Related Works

A. Direct S2UT

Lee et al. were the first to propose a discrete unit-based S2ST system, utilizing a Transformer Encoder-Decoder structure to directly establish the relationship between the source mel spectrogram and the target discrete units[6]. This structure is very similar to that of S2T and employs the intermediate inputs of the encoder to conduct a multitask auxiliary recognition. They used HuBert to extract speech representations of the target language, followed by k-means clustering to derive the target language's discrete units. Leveraging the self-supervised learning's characteristic of not requiring paired text, they also proposed S2ST for unwritten languages[11], [12], which similarly only required the target language's discrete units rather than text. To address the issue of varying units for different speakers, Lee et al. further introduced norm-unit-based S2ST, which was applied to textless real data, thereby enhancing the robustness and applicability of the system[7]. Popuri et al. used pre-trained wav2vec instead of a transformer encoder to extract acoustic features, improving the model's effect[13]. Therefore, in our research, we also apply the S2UT method using pretrained wav2vec as the encoder. We will introduce the structure of our S2UT in the next section.

B. Speaker Information Retention

Preserving speaker-specific information during speech translation has been a long-standing challenge. Recent advancements have leveraged techniques from speaker verification to enhance multi-speaker speech synthesis. Jia et al. integrated a speaker encoder to extract speaker features then combined with the synthesizer encoder, facilitating the retention of speaker identity in the synthesized speech[14]. Translatotron also employed a similar approach by adding a speaker encoder to extract speaker information[2]. Therefore, we adopted this structure as well, with the difference being that we replaced the construction of the speaker encoder. Our method utilizes an ECAPA-TDNN-based speaker encoder, which has demonstrated superior performance in speaker verification tasks[15]. The detailed architecture of the ECAPA-TDNN will be elaborated in the next section.

III. Speaker-Consistent Speech-to-Speech Translation

A. Model Structure

Fig.1 illustrates the overall architecture of our proposed SC-S2ST. From left to right, it shows the overall framework, the SR-U2M structure, and the speaker adapter structure within SR-U2M. The source language speech is input into the S2UT module, which generates a sequence of speech units in the target language. These units are then processed by the Length Predictor to match the actual sequence. Simultaneously, the source speech is input into the Speaker Adapter to extract speaker-specific information. The SR-U2M module integrates the speech units and the extracted speaker information to generate a mel-spectrogram of the target speech, retaining the original speaker's characteristics. Finally, a mel vocoder converts the mel-spectrogram into the waveform of the target speech.

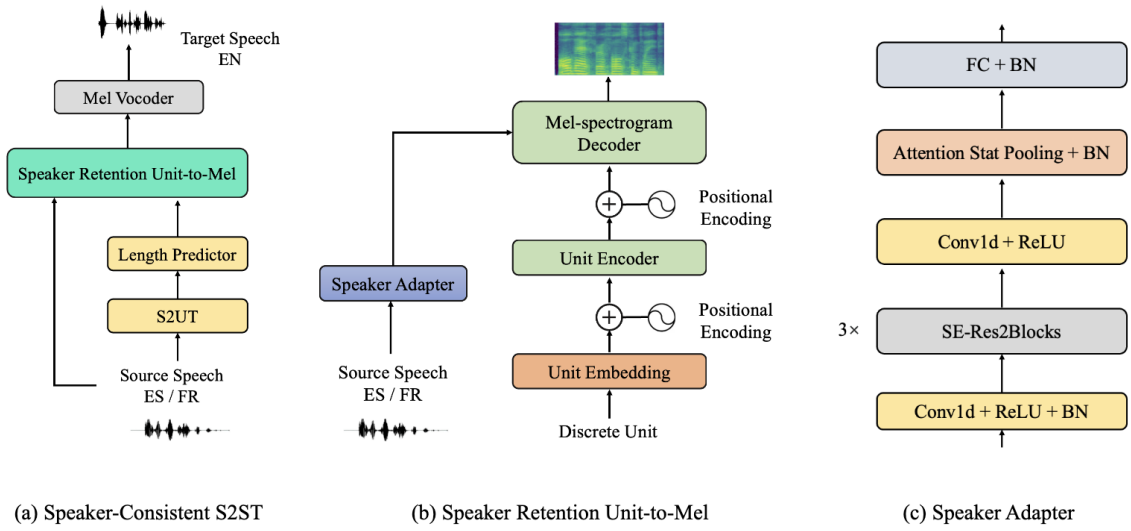
B. S2UT Model

Different from previous S2UT literature that used X-former (e.g., Transformer, Conformer)[6], [7], [11], [12], we utilize a pre-trained wav2vec2 model as the encoder. Wav2vec2 can extract richer speech representations from the original audio. For the decoder, we use the same Transformer decoder as others. Before training, we use the pre-trained HuBERT model to extract speech representations of the target language and then apply k-means clustering to obtain discrete units. Following Lee et al. we use a stacking method to convert adjacent identical units into a single unit, significantly speeding up both training and inference times[6]. Therefore, after S2UT, we need to add a length predictor to restore the units. Our length predictor is also derived from FastSpeech2, which consists of two 1D-convolutional layers, each with ReLU activation, followed by layer normalization and dropout, and a linear layer[16].

C. Speaker-Retention U2M

Our unit-to-mel approach is inspired by text-to-speech (TTS) systems, which convert discrete representations such as phonemes and text into mel-spectrograms. Examples of such systems include Tacotron and its variants[17], [18]. However, Tacotron uses an auto-regressive method, which is time-consuming during inference. To address this issue, we chose a FastSpeech-like structure that uses a non-auto-regressive approach to convert units into mel-spectrograms.

For the unit encoder, we employ the Feed-Forward Transformer (FFT) structure from FastSpeech[10]. This structure is similar to the encoder in the Transformer[19] model but



(a) Speaker-Consistent S2ST (b) Speaker Retention Unit-to-Mel (c) Speaker Adapter

Fig. 1. The overall architecture for Speaker-Consistent Speech-to-Speech Translation.

uses a 2-layer 1D convolutional network with ReLU activation instead of a 2-layer dense network, which has been shown to be more effective. Our mel-spectrogram decoder employs the traditional Transformer decoder structure. The specific parameters and configurations of our model will be detailed in the next section.

D. Speaker Adapter

The speaker adapter in our system is responsible for extracting speaker-specific information while disregarding other content information. Achieving this task primarily involves speaker identification[20]. The speaker adapter is designed to isolate and capture speaker-specific characteristics from the input speech signal. Our speaker adapter uses the ECAPA-TDNN architecture in its entirety to extract high-level speaker embeddings[15]. The structure is shown in subfigure(c) in Fig.1, which begins with a 1-dimensional convolutional layer followed by ReLU activation and batch normalization, capturing temporal features from the input speech signal. This is followed by three SE-Res2Blocks, which integrate Squeeze-and-Excitation mechanisms within residual blocks to emphasize important speaker-specific features. Another 1-dimensional convolutional layer with ReLU activation further refines these features. Attention-based statistical pooling then aggregates the most relevant frames into a fixed-dimensional speaker embedding, normalized by batch normalization. The final fully connected layer, also followed by batch normalization, transforms these aggregated features into the final speaker embedding, effectively encapsulating the unique vocal characteristics of the speaker.

IV. Experiments

A. Dataset

We performed our experiments using Spanish to English and France to English in CVSS multilingual speech-to-speech translation corpus[21]. The CVSS corpus includes two datasets: CVSS-C, which contains multi-speaker recordings in various languages translated into the voice of a single female English speaker, and CVSS-T, which contains multi-speaker recordings translated into multi-speaker English speech, maintaining similar pronunciation characteristics across both datasets. The number of samples and their respective durations are detailed in TABLE I.

We utilize the CVSS-T corpus to train our SR-U2M model because it provides the necessary multi-speaker target speech mel-spectrogram required for our approach. Conversely, the CVSS-C corpus is employed to train the S2UT model. Since discrete units only encapsulate semantic information, extracting these units from the same speaker's voice, as provided in the CVSS-C corpus, yields better performance.

B. System Implementation

1) SC-S2UT Model: we used the pre-trained HuBERT model[1] trained on LibriSpeech[22] and follow illustrate k-means with $K = 100$ to extract sixth layers representation to discrete unit. We utilized the pre-trained wav2vec to extract acoustic feature from source waveform, which trained on the es unlabeled subset of VoxPopuli corpus[23] for Spanish[2], and trained on 14K hours of French[3][24]. We did not employ multi-task learning in our experiments as focus of our study is to compare the extent to which our method can retain

[1]https://github.com/facebookresearch/fairseq/tree/main/examples/hubert
[2]https://github.com/facebookresearch/voxpopuli/
[3]https://huggingface.co/LeBenchmark/wav2vec2-FR-14K-large

TABLE I
STATISTICS (NUMBER OF SAMPLES AND DURATION) OF THE CVSS SPANISH-ENGLISH AND FRANCE-ENGLISH DATASET

	CVSS-C								CVSS-T		
	Es-En				Fr-En				Es-En		
	train	dev	dev-small	test	train	dev	dev-small	test	train	dev	test
# samples	79k	13.2k	0.99k	13.2k	20.7k	14.7k	0.49k	14.7k	79k	13.2k	13.2k
source (hrs)	69.5	12.4	1.31	12.4	174	13	0.64	13.3	73.7	13.2	13.3
target (hrs)	113.1	21.8	/	22.7	264.3	21.7	/	23.3	113.1	21.8	22.7

speaker information during the translation process. For SR-U2M, we compute 80-dimensional mel-spectrogram features at every 20-ms, making it matching the length of discrete unit. Hyperparameters of our model is shown in TABLE II.

TABLE II
HYPERPARAMETERS FOR SC-S2ST

Hyperparameter		SC-S2ST
S2UT Decoder	Unit neurons	103
	Decoder Block	6
	Decoder Hidden	512
	Encoder Attention Heads	8
	Encoder Dropout	0.1
SR-U2M Unit-Encoder Mel-Decoder	Encoder Block	6
	Decoder Block	6
	Encoder Kernel	31
	Hidden	512
	Attention Heads	8
	Dropout	0.1
Speaker Adapter	Channels	[1024, 1024, 1024, 1024, 3072]
	Kernel Sizes	[5, 3, 3, 3, 1]
	Dilations	[1, 2, 3, 4, 1]
	Groups	[1, 1, 1, 1, 1]
	Attention Channels	128

For waveform generation, we used the same unit-based vocoder as [13] serving as a baseline for comparison. In our SC-S2UT system, since the final output is a mel-spectrogram, we employ HiFi-GAN[5] as the vocoder[4] and trained with multispeaker LibriTTS corpus[25].

2) Baselines: We developed two cascaded baselines, ASR+MT+TTS and S2T+TTS, and one direct S2ST baseline, which also utilized a pre-trained wav2vec2 model as the encoder. For the Automatic Speech Recognition (ASR) component, we employed wav2vec2 pre-trained on XLSR[26] and fine-tuned it with Connectionist Temporal Classification (CTC) for French[5] and Spanish[6] speech recognition. For Machine Translation (MT), we used a pre-trained transformer model for both French and Spanish[7]. For Text-to-Speech (TTS), we leveraged the Massively Multilingual Speech (MMS) model for English text-to-speech[8][27]. Additionally, we used a transformer-based sequence-to-sequence (seq2seq) model for Spanish and French speech-to-text translation[28]. Furthermore, we implemented a wav2vec-based S2UT[13] and used a Unit-HiFi-GAN vocoder to convert the units into waveforms.

3) Evaluation: We evaluated the system's performance in three aspects: translation quality, speech quality, and speaker similarity. For translation quality, we followed previous literature by applying ASR to the speech output and computing BLEU scores. We used a transformer-based end-to-end model pre-trained on LibriSpeech, which achieved a WER of 2.27 on the test-clean set[9]. For translation quality, we used all development and test data in Spanish and French. For speech quality evaluation, we used only test data and utilized Mean Opinion Scores (MOS) to assess the naturalness of the output speech, with a scale from 1 (worst) to 5 (best). We used UTMOS based on the methodology described in [29]. We also conducted speed tests using the baseline for comparison. Our experiments were performed on an RTX 3090 GPU. Lastly, for speaker similarity evaluation, we used the CVSS-T corpus, which features source and target speakers with similar voices. Additionally, because our SR-UTM structure was trained using units and target speech, we compared the similarity between the translated speech and target speech as well as source speech. We extracted speaker features from the output speech using the Speaker Adapter structure within our system shown in subfigure(c) in Fig.1. We then computed normalized Mean Squared Error (NMSE), Standard Euclidean Distance (SED), and Cosine Similarity (COS-SIM) between different features to quantify speaker similarity. Since we trained our SR-UTM using only Spanish data, for similarity evaluation, we used a Spanish-English dataset. All our experiments were conducted using the SpeechBrain framework[30].

V. RESULTS

In this section, we present the results of our experiments on translation quality and speech quality. The primary goal of our approach is to maintain the source speaker's voice while achieving competitive translation and speech quality.

A. Translation Quality and Speech Quality

TABLE III compares the BLEU scores and MOS scores and inference time and speedup of our proposed method against the

[4]https://huggingface.co/speechbrain/tts-hifigan-libritts-16kHz
[5]https://huggingface.co/facebook/wav2vec2-large-xlsr-53-french
[6]https://huggingface.co/facebook/wav2vec2-large-xlsr-53-spanish
[7]https://github.com/Helsinki-NLP/Tatoeba-Challenge/tree/master/models
[8]https://huggingface.co/facebook/mms-tts-eng
[9]https://huggingface.co/speechbrain/asr-transformer-transformerlm-librispeech

TABLE III
BLEU SCORE AND MOS SCORE AND INFERENCE TIME AND SPEEDUP FOR OUR EXPERIMENT

	BLEU Score ↑				MOS Score ↑		Inference Time(s) ↓		Speedup ↑	
	ES-EN		FR-EN		ES-EN	FR-EN	ES-EN	FR-EN	ES-EN	FR-EN
	dev	test	dev	test	test	test	dev-small	dev-small	test	test
Direct System										
S2UT[13]	15.56	16.54	22.26	22.67	3.31 ± 0.11	3.29 ± 0.12	0.77	0.82	1.00×	1.00×
SC-S2UT	15.03	16.10	21.52	21.68	3.26 ± 0.13	3.20 ± 0.11	1.00	1.09	0.77×	0.75×
Cascade System										
ASR+MT+TTS	22.16	24.36	23.36	23.22	4.07 ± 0.06	4.10 ± 0.08	/	/	/	/
ST+TTS	13.82	14.13	21.05	20.72	4.11 ± 0.08	4.10 ± 0.07	/	/	/	/
Ground Truth	84.51	88.64	80.51	80.29	4.45 ± 0.11	4.44 ± 0.11	4.74	4.75	/	/

baseline. The MOS score is presented with the standard deviation. From the results, we observed that the ASR+MT+TTS cascade system still achieved the highest BLEU scores in both language pairs and datasets. Our proposed SC-S2UT method showed slightly lower BLEU scores compared to the baseline S2UT, indicating a small trade-off in translation quality when focusing on preserving the speaker's voice. However, the gap between our method and S2UT is very small, demonstrating that while our approach might lose a bit of content information due to its emphasis on retaining speaker characteristics, the loss is minimal and does not significantly impact content comprehension. In TableIII, inference time represents the average duration of inference, while ground truth denotes the average duration of the source speech. The results show that our method only slightly increases the required inference time, reducing the speed to approximately 0.76 times that of the original. Our proposed SC-S2UT method, while having a slightly lower Mean Opinion Score (MOS) compared to the cascade systems, is comparable to the baseline S2UT, achieving near state-of-the-art speech quality without significantly increasing inference time, thereby maintaining efficiency. Overall, our method, with the goal of preserving speaker information, achieved results very close to the baseline S2UT system.

B. Speaker Similarity

For speaker similarity evaluation, we compared the similarity between the speech obtained from the translation of source speech with the source speech and target speech, as shown in TABLE IV.

Based on the experimental results, we can find that, first, it is expected that the "*tal-tgt*" scenario shows the highest similarity, reflected by the COS-SIM score of 0.84 on the test set, and the lowest NMSE (0.051) and SED (0.51) values. These results indicate that our model effectively captures the characteristics of the target speech, as it was trained using target speech data. More importantly, the comparison between the translation and source speech demonstrates higher similarity and better performance in terms of NMSE and SED compared to the direct comparison between source and target speeches. Specifically, the COS-SIM score for "*tal-src*" (0.79) is higher than that for "*src-tgt*" (0.71). Additionally, the NMSE and SED values for "*tal-src*" are lower than those for "*src-tgt*" suggesting that our translation process preserves the speaker-specific information from the source speech more effectively than the direct reference between source and target. These findings suggest that our model excels in maintaining speaker similarity and preserving crucial speaker-specific attributes from the source speech.

TABLE IV
SPEAKER SIMILARITY EXPERIMENT RESULTS

	NMSE↓		SED↓		COS-SIM↑	
Compared Item	dev	test	dev	test	dev	test
tal-src	**0.070**	0.071	**0.60**	0.60	0.79	0.79
tal-tgt	0.075	**0.051**	0.64	**0.51**	**0.84**	**0.82**
src-tgt	0.089	0.072	0.71	0.62	0.77	0.71

"*tal-src*" refers to the comparison between translation and source speech, "*tal-tgt*" denotes the comparison between translation and target speech, and "*src-tgt*" represents the comparison between source speech and target speech. (↑) indicates that higher values are better, while (↓) mens that lower values are better.

VI. CONCLUSION

In this paper, we proposed a SR-UTM architecture designed to preserve the voice characteristics of the source speaker during speech-to-speech translation with low inference latency. We conducted experiments on the CVSS-C and CVSS-T corpora. The results demonstrated that our SC-S2UT system achieved translation quality comparable to the baseline S2UT system. Moreover, the speaker similarity experiments confirmed that our method effectively retains the speaker-specific information, without significantly increasing inference time. This highlights the robustness and efficacy of our approach in maintaining speaker identity during translation.

REFERENCES

[1] A. Bérard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," *arXiv preprint arXiv:1612.01744*, 2016.

[2] Y. Jia, R. J. Weiss, F. Biadsy, *et al.*, "Direct speech-to-speech translation with a sequence-to-sequence model," *arXiv preprint arXiv:1904.06037*, 2019.

[3] Y. Jia, M. T. Ramanovich, T. Remez, and R. Pomerantz, "Translatotron 2: Robust direct speech-to-speech translation," 2021.

[4] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[5] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[6] A. Lee, P.-J. Chen, C. Wang, *et al.*, "Direct speech-to-speech translation with discrete units," *arXiv preprint arXiv:2107.05604*, 2021.

[7] A. Lee, H. Gong, P.-A. Duquenne, *et al.*, "Textless speech-to-speech translation on real data," *arXiv preprint arXiv:2112.08352*, 2021.

[8] K. Song, Y. Ren, Y. Lei, *et al.*, "Styles2st: Zero-shot style transfer for direct speech-to-speech translation," *arXiv preprint arXiv:2305.17732*, 2023.

[9] Y. Wang, J. Bai, R. Huang, R. Li, Z. Hong, and Z. Zhao, "Speech-to-speech translation with discrete-unit-based style transfer," *arXiv preprint arXiv:2309.07566*, 2023.

[10] Y. Ren, Y. Ruan, X. Tan, *et al.*, "Fastspeech: Fast, robust and controllable text to speech," *Advances in neural information processing systems*, vol. 32, 2019.

[11] C. Zhang, X. Tan, Y. Ren, T. Qin, K. Zhang, and T.-Y. Liu, "Uwspeech: Speech to speech translation for unwritten languages," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 14 319–14 327.

[12] P.-J. Chen, K. Tran, Y. Yang, *et al.*, "Speech-to-speech translation for a real-world unwritten language," *arXiv preprint arXiv:2211.06474*, 2022.

[13] S. Popuri, P.-J. Chen, C. Wang, *et al.*, "Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation," *arXiv preprint arXiv:2204.02967*, 2022.

[14] Y. Jia, Y. Zhang, R. Weiss, *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in neural information processing systems*, vol. 31, 2018.

[15] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020*, 2020, pp. 3830–3834.

[16] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[17] Y. Wang, R. Skerry-Ryan, D. Stanton, *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[18] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 4779–4783.

[19] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," *Expert Systems with Applications*, vol. 90, pp. 250–271, 2017.

[21] Y. Jia, M. T. Ramanovich, Q. Wang, and H. Zen, "Cvss corpus and massively multilingual speech-to-speech translation," *arXiv preprint arXiv:2201.03713*, 2022.

[22] V. Panayotov, G. Chen, D. Povey, and S. 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.

[23] C. Wang, M. Riviere, A. Lee, *et al.*, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," *arXiv preprint arXiv:2101.00390*, 2021.

[24] T. Parcollet, H. Nguyen, S. Evain, *et al.*, "Lebenchmark 2.0: A standardized, replicable and enhanced framework for self-supervised representations of french speech," *Computer Speech & Language*, vol. 86, p. 101 622, 2024.

[25] H. Zen, V. Dang, R. Clark, *et al.*, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[26] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[27] V. Pratap, A. Tjandra, B. Shi, *et al.*, "Scaling speech technology to 1,000+ languages," *arXiv*, 2023.

[28] C. Wang, Y. Tang, X. Ma, *et al.*, "Fairseq s2t: Fast speech-to-text modeling with fairseq," *arXiv preprint arXiv:2010.05171*, 2020.

[29] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "Utmos: Utokyo-sarulab system for voicemos challenge 2022," *arXiv preprint arXiv:2204.02152*, 2022.

[30] M. Ravanelli, T. Parcollet, P. Plantinga, *et al.*, *SpeechBrain: A general-purpose speech toolkit*, arXiv:2106.04624, 2021. arXiv: 2106 . 04624 [eess.AS].