Improving Speech-to-Speech Translation for Low-Resource Languages via Transfer Learning

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

***Abstract*—End-to-end Speech-to-Speech Translation (S2ST) enables direct transformation from source speech in one language to target speech in another. Still, it typically requires large-scale parallel data, which is unavailable for many low-resource languages. In this work, we investigate transfer learning strategies to improve S2ST performance in low-resource settings. Our approach involves pretraining a multi-task S2ST model on high-resource language pairs (e.g., French-English, German-English, Spanish-English) and transferring parameters to models for low-resource pairs such as Italian-English. We evaluate multiple transfer configurations, including full S2UT transfer, auxiliary ASR module sharing, and partial encoder/decoder reuse. Experimental results demonstrate that transfer from French-English consistently yields the most significant gains, owing to its larger data volume. BLEU scores improve from 10.41 to 13.77 (ES-EN), 9.70 to 12.15 (DE-EN), and more notably from 2.89 to 10.01 and 2.10 to 9.10 for the low-resource IT-EN and RU-EN pairs. Ablation studies show that encoder transfer contributes more to performance than decoder transfer, and that auxiliary ASR modules provide limited benefits. These findings suggest that carefully designed structural transfer, particularly of the encoder, is crucial for effective low-resource S2ST.**

I. Introduction

Speech-to-Speech Translation (S2ST) enables direct conversion of spoken utterances from one language to another, significantly reducing barriers in cross-lingual communication. In recent years, end-to-end S2ST models have gained attention for their simplicity and reduced inference latency compared to traditional cascade systems composed of Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS) modules.

The first end-to-end S2ST model, Translatotron, employed an encoder-decoder architecture to directly map the source speech's melspectrogram to the target melspectrogram. To improve the model's representational capacity, it incorporated auxiliary tasks in a multi-task learning framework[1]. Its successor, Translatotron 2, introduced a two-stage optimization strategy, where the speech translation decoder contributes during inference, further enhancing performance[2].

Another paradigm is Speech-to-Unit Translation (S2UT), which first uses a self-supervised model (e.g., HuBERT[3]) trained on the target language to extract speech representation. These representations are then clustered into discrete units, allowing the source speech to be translated into target units[4]. A vocoder, such as Unit HiFi-GAN, is subsequently used to synthesize the final waveform[5].

Despite these advances, training high-quality S2ST systems requires large amounts of parallel speech data, which are scarce for most language pairs. S2ST remains a major challenge under such low-resource conditions. In contrast, low-resource speech recognition has achieved notable success through transfer learning, where knowledge from a high-resource domain is leveraged to enhance performance in a related low-resource domain. Since speech across languages shares common acoustic characteristics being ultimately produced by human vocal tracts, this makes transfer learning a natural fit for multilingual speech modeling.

In this work, we explore the use of transfer learning to improve S2ST in low-resource language settings. Specifically, we pretrain multi-task S2ST models on high-resource language pairs and transfer the learned parameters to low-resource directions. Furthermore, we conduct a detailed ablation study to examine the effectiveness of different transferred components (e.g., encoder, decoder, and auxiliary ASR modules), revealing which parts are most crucial for successful adaptation. Our audio samples are available at[1].

II. Related research

A. Direct S2UT System

In this work, we adopt an S2UT architecture, which has emerged as an effective alternative to Transformer-based end-to-end S2ST systems that use a melspectrogram. The original S2UT model employs a Transformer-based encoder-decoder architecture to autoregressively generate discrete units in the target language directly from source speech[4].

Several extensions to the original S2UT framework have since been proposed. For example, Lee et al. pointed out that the same utterance spoken by different speakers often yields distinct unit sequences due to speaker dependent acoustic variations. To address this, they introduced unit normalization, which reduces speaker-induced variance and improves translation quality[6]. Huang et al. proposed TransSpeech which used a bilateral perturbation technique to suppress speaker-specific factors such as timbre and energy, retaining only the content relevant components to enhance unit prediction[7]. UnitY extended the S2UT paradigm into a two-stage model inspired by Translatotron 2. The source speech is first transcribed into target text, which is then converted into discrete units, resulting in improved performance over single-stage models[8].

In our study, we focus on analyzing how transfer learning affects the performance of S2UT in low-resource settings. Due

[1]https://zhouruitohoku99.github.io/transfers2ut/

979-8-3315-7206-8/25/$31.00

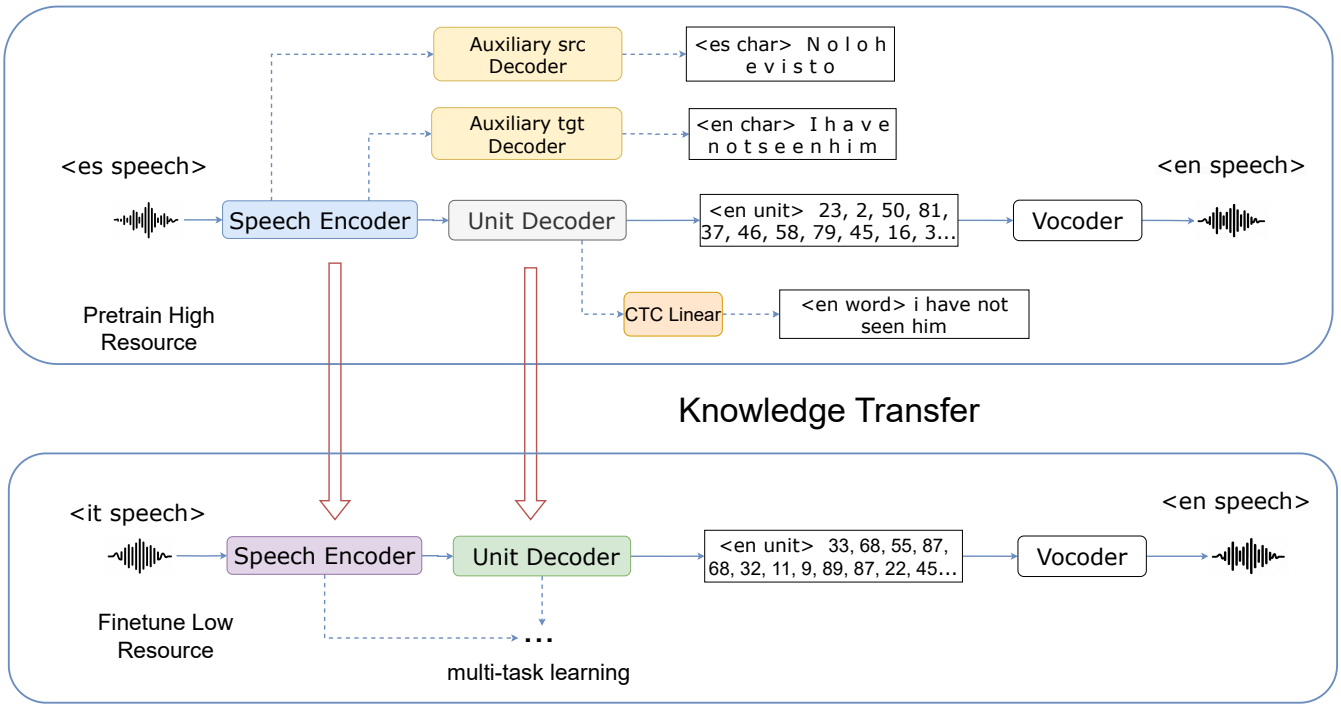

Fig. 1. Architecture of our transfer learning approach for end-to-end speech-to-speech translation. The upper part shows the multi-task pretraining on high-resource language pairs, and the lower part shows fine-tuning on low-resource language pairs.

to data limitations, we base our experiments on the original single stage S2UT system[4], without incorporating additional intermediate text or normalization modules.

B. Transfer Learning in Speech Processing

Owing to the inherent acoustic similarities across different human languages, transfer learning has become a widely adopted strategy in low-resource speech processing tasks. The core idea is to transfer representations or model parameters learned from high-resource conditions to low-resource settings, enabling models to achieve better performance with limited labeled data.

In the field of speech recognition, several studies have successfully applied transfer learning. For example, Yu et al. transferred knowledge learned from a Mandarin corpus to improve ASR performance on Tujia[9], a low resource Chinese dialect. Similarly, Tong et al. leveraged adult speech to enhance ASR accuracy for children, demonstrating effective cross domain adaptation[10]. Zhou et al. proposed meta transfer learning for multi low resource language speech recognition[11]. Hsu et al. proposed a novel transfer learning methods, in which robust phonetic features are extracted from grounding models based on the semantic correlation between images and speech, without using transcripts [12].

Transfer learning has also been widely adopted in speech synthesis, especially to address the challenges of speaker adaptation and low resource settings. Jia et al. incorporated speaker embeddings learned from a speaker verification model into a multi speaker TTS system, enabling the generation of diverse and speaker consistent speech[13]. Others applied cross-lingual transfer learning to end-to-end TTS, demonstrating that knowledge from high-resource languages can benefit low-resource TTS models[14], [15].

These works collectively demonstrate that cross domain and cross-lingual transfer can be highly effective in both recognition and synthesis tasks. Inspired by these findings, our work applies structural transfer learning to speech-to-speech translation a task that combines both recognition and synthesis and analyzes the effectiveness of different transfer configurations, particularly in low resource scenarios.

III. Methods

In this section, we describe our proposed method for improving S2ST in low-resource scenarios through cross-lingual transfer learning.

A. Overview

Fig. 1 illustrates the overall architecture of our proposed system. Our system is built upon the S2UT framework and trained using a multi-task learning strategy to enhance the quality of encoder representations. In the pretraining stage on hig-resource language pairs, the input source speech is first processed by a Transformer-based encoder. The encoder outputs are then used to optimize three objectives simultaneously. First, a unit decoder autoregressively generates a sequence of discrete target units. Second, an auxiliary decoder receives intermediate encoder representations to predict the characters in the source language. Third, another auxiliary decoder takes

TABLE I
STATISTICS OF THE CVSS-C CORPUS USED IN OUR EXPERIMENTS. EACH LANGUAGE PAIR CONTAINS NATURALLY SPOKEN SOURCE SPEECH AND SYNTHETIC ENGLISH SPEECH AS TARGET.

	ES-EN			FR-EN			DE-EN			IT-EN			RU-EN		
	Train	**Dev**	**Test**	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**
# Samples	79k	13.2k	13.2k	207k	14.7k	14.7k	127.8k	13.5k	13.5k	31.6k	8.9k	8.9k	12.1k	6.1k	6.3k
Source hours	113.1	21.8	22.7	264.3	21.7	23.3	184.3.	20.5	21.5	44.2	14.3	15.4	18.2	9.9	10.6
Target hours	69.5	12.4	12.4	174	13	13.3	112.4	12.5	12.1	29.4	8.5	8.6	13.3	6.7	6.9

deeper encoder outputs to predict target language characters. In addition, we perform target word recognition by applying a CTC loss to the intermediate hidden states of the unit decoder. The two auxiliary multi-task learning are only involved during training, as indicated by the dashed arrows in the figure, and are omitted during inference.

B. Knowledge Transfer and Fine-Tuning

After pretraining on high-resource language pairs, we transfer the learned parameters to initialize a new model for low-resource language speech-to-speech translation. The goal is to leverage the representations learned from large-scale data to improve performance in scenarios where training data is scarce. In our framework, we reuse the parameters of the speech encoder and unit decoder, which are shared across all tasks during pretraining. The auxiliary decoders for source side and target side ASR, as well as the intermediate decoder CTC module for target word recognition, may also be transferred depending on the configuration.

To better understand the role of each component in transfer learning, we design several configurations with varying degrees of parameter reuse. These include full model transfer, where all encoder, decoder, and auxiliary components are transferred; partial transfer, where only the encoder or specific layers of the decoder are reused; and minimal transfer, where only the core S2UT model (encoder and unit decoder) is retained. All models are then fine-tuned on low-resource language pairs using the same multi-task learning objectives as in the pretraining phase. Through this process, we aim to analyze which parts of the model contribute most to effective adaptation in low-resource S2ST scenarios.

IV. EXPERIMENTS

A. Dataset

We conduct our experiments using the CVSS-C corpus, a publicly available multilingual speech-to-speech translation dataset. CVSS-C consists of speech data from multiple speakers in various source languages paired with automatically generated English target speech[16]. The English speech is synthesized using a single speaker female TTS system, while the source speech is naturally spoken and includes a diverse range of speaker identities, accents, and prosody patterns.

The corpus covers several language pairs, among which we select three high-resource languages pairs: French-English (FR-EN), Spanish-English (ES-EN), and German-English (DE-EN) and two relatively low-resource languages pairs Italian-English (IT-EN) and Russian-English (RU-EN) for evaluation. During the pretraining stage, we use the FR-EN, ES-EN, and DE-EN pairs to train our base models. These pretrained models are then fine-tuned on different language directions in the adaptation stage to simulate cross-lingual transfer in low-resource conditions.

The statistics of the training and validation sets for each language pair used in our experiments are summarized in TABLE I.

B. System Implementation

Our model follows the overall architecture of S2UT, proposed in [4], which consists of a transformer-based encoder and decoder combined with a multi-task learning setup. The speech encoder comprises 12 Transformer encoder layers, and the unit decoder consists of 6 Transformer decoder layers. We also utilize auxiliary recognition tasks using intermediate encoder outputs. Specifically, the output of the 6th encoder layer is fed into an auxiliary transformer-based source character recognition decoder, which consists of two additional decoder layers. Similarly, the 8th encoder layer output is used by a separate target character recognition decoder, also implemented as a 2-layer Transformer decoder. In addition to the auxiliary ASR tasks, we incorporate a third multi-task objective into the unit decoder. The output of the 3rd decoder layer is passed through a linear projection and optimized using Connectionist Temporal Classification loss to predict the target word sequence.

For unit generation, we follow the standard speech-to-unit pipeline by extracting speech representations from the 6th layer of a pretrained HuBERT model[2], and clustering them using K-means with K = 100 to obtain discrete unit tokens. We set the vocabulary size of source and target character tokens to 50, and define a target word vocabulary of size 1000, constructed using a unigram language model.

To ensure consistent structure across all language pairs, we apply the same model architecture and hyperparameter settings for both pretraining and finetuning. This ensures compatibility during cross-lingual transfer and minimizes variance resulting from architectural mismatch.

[2]https://github.com/facebookresearch/fairseq/tree/main/examples/hubert

TABLE II
BLEU SCORES ON VALIDATION AND TEST SETS ACROSS DIFFERENT LANGUAGE PAIRS AND SYSTEM CONFIGURATIONS.

Method	ES-EN		DE-EN		FR-EN		RU-EN		IT-EN	
	Valid	Test	Valid	Test	Valid	Test	Valid	Test	Valid	Test
Cascade System										
ASR[17] + MT[18] + TTS[19]	22.16	24.36	21.57	20.98	23.36	23.22	31.44	30.89	14.76	14.89
ST[20] + TTS[19]	13.83	14.13	10.23	10.84	21.05	20.72	-	-	-	-
Direct System										
S2UT[4]	10.41	11.10	9.70	9.32	17.02	16.48	2.10	2.06	2.89	2.77
S2UT + Pretrained on ES-EN	-	-	11.05	10.47	**17.21**	**16.77**	7.57	8.18	8.17	8.03
S2UT + Pretrained on DE-EN	11.85	12.92	-	-	17.13	16.56	8.19	8.94	7.07	7.11
S2UT + Pretrained on FR-EN	**13.77**	**14.75**	**12.15**	**11.66**	-	-	**9.10**	**10.11**	**10.01**	**9.92**
Ground Truth	84.51	88.54	82.32	81.09	80.51	80.29	84.39	85.22	81.15	82.08

C. Baselines and Evaluation Metrics

To evaluate the effectiveness of our proposed transfer learning framework, we compare it with several baseline systems, including traditional cascade approaches and a non-transfer version of the S2UT model. We consider two cascade systems: (1) an ASR→MT→TTS pipeline and (2) an ST→TTS pipeline. In the first system, the input speech is transcribed using the multilingual wav2vec 2.0 large model (XLS-R-53)[3][17], followed by machine translation with Opus-MT[4][18], and finally converted into speech using the English TTS model from the Massively Multilingual Speech (MMS) project[5][19]. In the second cascade system, we use a speech translation model pretrained on CoVoST2 to directly translate speech into English text[20], which is then synthesized using the same MMS English TTS model. As an additional baseline, we include a direct S2UT model trained without any parameter transfer.

All systems in our experiments generate English speech as the target output. To evaluate the semantic accuracy of the generated speech, we first transcribe the synthesized English speech using a Transformer-based ASR model pretrained on LibriSpeech, which achieves a word error rate (WER) of 2.27% on the LibriSpeech test-clean set[6]. The resulting transcriptions are then compared against reference translations using case-insensitive detokenized BLEU scores, computed with the SacreBLEU toolkit[21].

V. RESULTS

A. Effectiveness of Transfer Learning

Table II presents the BLEU scores on validation and test sets across multiple language pairs, focusing on the effectiveness of transfer learning in the direct S2UT architecture. We observe that models without any pretraining consistently underperform across all language directions, especially under low resource settings such as RU-EN and IT-EN, where BLEU scores remain close to unusable. In contrast, incorporating transfer learning with high resource language pairs significantly improves translation quality.

Among all configurations, pretraining on FR-EN yields the most substantial improvements across all target language pairs. This can be largely attributed to the abundance of FR-EN training data, which allows the model to learn more robust mappings between speech and discrete unit sequences. For instance, the BLEU score on IT-EN improves from 2.77 to 9.92 with FR-EN pretraining, a more than threefold increase.

Beyond data quantity, linguistic similarity between the source and target languages also plays a vital role in transfer effectiveness. As French, Spanish, and Italian all belong to the Romance language family, using ES-EN as a pretraining source leads to better results on IT-EN compared to DE-EN, which is from a different language family (8.03 vs. 7.11). Conversely, for RU-EN, which is linguistically distant from both ES and FR, pretraining on DE-EN is more effective than on ES-EN (8.94 vs. 8.18), highlighting the importance of structural compatibility in cross-lingual transfer.

From these findings, we draw two key conclusions. First, the effectiveness of transfer learning is highly dependent on the amount of data available for the source language pair. Second, linguistic proximity between the source and target languages significantly influences the success of transfer, with closer languages yielding better adaptation performance.

Nevertheless, it is worth noting that a performance gap remains between our direct S2UT models and the cascade systems. This discrepancy may largely stem from the limited availability of parallel speech data for direct training.

B. Results by Transfer Module Configuration

Fig.2 presents the BLEU scores obtained under different module-level transfer strategies for S2ST. The upper subplot corresponds to a low-resource language pair IT-EN, while the lower subplot shows results for a high-resource pair ES-EN. All of them were pretrained by FR-EN.

[3]https://huggingface.co/facebook/wav2vec2-large-xlsr-53
[4]https://github.com/Helsinki-NLP/Opus-MT
[5]https://huggingface.co/facebook/mms-tts-eng
[6]https://huggingface.co/speechbrain/asr-transformer-transformerlm-librispeech

TABLE III

QUALITATIVE COMPARISON OF GENERATED TRANSLATIONS UNDER DIFFERENT TRANSFER SETTINGS ON IT-EN. EACH BLOCK SHOWS TRANSLATIONS WITHOUT PRETRAINING, WITH DECODER-ONLY OR ENCODER-ONLY TRANSFER, AND WITH FULL MODEL TRANSFER.

Source:	Fu la prima donna sepolta nel cimitero.
Reference:	she was the first woman to be buried in the cemetery
No Pretrain:	he **was the first woman** had been in the sea ground
Transfer Decoder:	he **was the first woman** with his first **cemetery**
Transfer Encoder:	it **was the first woman** varied **in the cemetery**
Transfer All:	it **was the first woman buried in the cemetery**
Source:	Queste competenze risultano indispensabili per l'attività di educatore che operi nell'ambito di un sistema formativo
Reference:	these competencies are indispensable for the educator activities who operates in the education system
No Pretrain:	suspension resigns and individual partic abilities that required of every difficulty in orderable
Transfer Decoder:	his six mansions results in experimental field that allows you to expanding the devices for an avity'
Transfer Encoder:	his **competence** is an essential training the respectivity of ducator that **operated** some feeling
Transfer All:	**these competency is inespensable** to incest the **activity that opposes** in the environment of training **system**

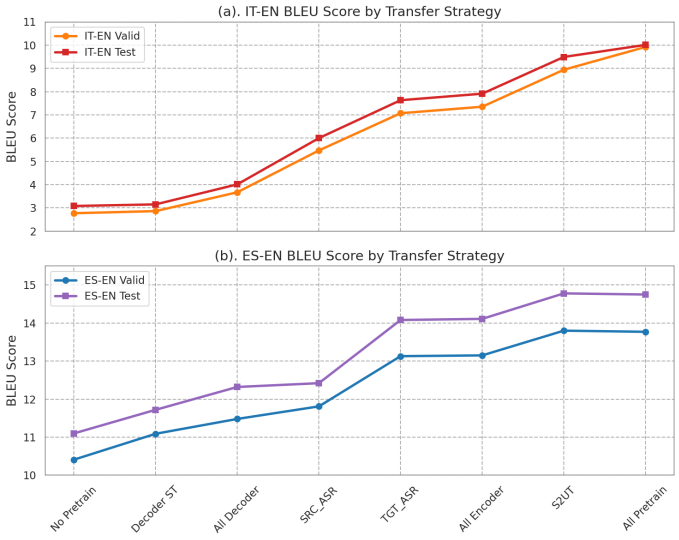

Fig. 2. BLEU score comparison of different module-level transfer strategies on IT-EN and ES-EN S2ST tasks. From left to right, the x-axis denotes progressively larger subsets of transferred modules: No Pretrain, Decoder ST (3 layer decoder + CTC), All Decoder, SRC_ASR (6 layer encoder + source character recognition decoder), TGT_ASR (8 layer encoder + target character recognition decoder), All Encoder, S2UT (encoder + decoder), and All Pretrain (entire model).

We first observe that transferring only the decoder either partially (Decoder ST) or entirely (All Decoder) yields only marginal improvements in both settings. This suggests that decoder parameters, in isolation, are not the primary contributors to cross-lingual generalization. Second, the performance gap between transferring TGT_ASR and All Encoder is minimal, indicating that transferring deeper encoder layers (e.g., layers 7–12) offers limited benefits. This underscores the importance of early encoder layers, which capture more language-agnostic acoustic representations, as the most transferable component across languages.

Third, in the high resource scenario, the performance difference between transferring All Encoder and transferring S2UT remains relatively small. However, in the low-resource setting, this gap widens significantly. This highlights that decoder transfer becomes increasingly valuable when adaptation data is limited, likely due to its role in guiding target side generation under scarce supervision. Lastly, incorporating auxiliary multi-task modules such as source or target side ASR decoders during transfer yields negligible gains, indicating that their utility primarily lies in enhancing pretraining, rather than in transferability.

Overall, these findings suggest that lower encoder layers are the most essential for effective transfer, and the benefit of decoder reuse scales with the level of resource scarcity in the target language.

C. Qualitative Evaluation on Real IT-EN Examples

We further conducted a qualitative analysis on real IT-EN test cases, as shown in Table III. We compared translations under four configurations: no pretraining, decoder-only transfer, encoder-only transfer, and full model transfer.

In shorter sentences, the model without pretraining was able to capture some surface-level structure but often failed to translate proper nouns and contextually grounded words. Decoder transfer led to minor improvements in recognizing certain terms but still lacked semantic coherence. Encoder transfer notably improved both semantic alignment and term recognition. Full model transfer achieved the most accurate translations, correctly rendering sentence structure and content.

In longer and more complex inputs, both no pretraining and decoder transfer produced incoherent outputs, failing to capture even basic semantics. Encoder transfer allowed the model to identify some key terms, but the translations remained fragmented. In contrast, full-model transfer enabled the model to extract most keywords and produce partially coherent sentences, reflecting a substantial improvement over the non-pretrained baseline even though some errors remained.

VI. CONCLUSIONS

This paper investigates transfer learning for end-to-end S2ST in low resource settings. By pretraining on high resource language pairs and fine-tuning on low-resource ones,

we achieved consistent performance gains, with FR-EN pretraining yielding the best results due to larger data volume. Ablation studies show that encoder transfer plays the most critical role, while decoder and auxiliary modules have limited impact. The full model transfer is especially beneficial in low-resource cases. Qualitative examples further confirm that transfer learning enhances both lexical accuracy and semantic coverage.

However, the performance gap between direct and cascade systems remains noticeable, particularly in low resource conditions. Future work will explore better unit representations, unsupervised pretraining methods, and parameter-efficient adaptation techniques to further close this gap and improve generalization to unseen language pairs.

Acknowledgment

This work was supported by JST SPRING, Grant Number JPMJSP2114.

References

[1] Y. Jia, R. J. Weiss, F. Biadsy, *et al.*, "Direct speech-to-speech translation with a sequence-to-sequence model," in *Proc. Interspeech 2019*, 2019, pp. 1123–1127.

[2] Y. Jia, M. T. Ramanovich, T. Remez, and R. Pomerantz, "Translatotron 2: Robust direct speech-to-speech translation," *arXiv preprint arXiv:2107.08661*, vol. 6, no. 7, 2021.

[3] 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.

[4] A. Lee, P.-J. Chen, C. Wang, *et al.*, "Direct speech-to-speech translation with discrete units," in *Proc. ACL, VOL.1*, 2022, pp. 3327–3339.

[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, H. Gong, P.-A. Duquenne, *et al.*, "Textless speech-to-speech translation on real data," in *Proc. NAACL-HLT*, 2022, pp. 860–872.

[7] R. Huang, J. Liu, H. Liu, *et al.*, "Transpeech: Speech-to-speech translation with bilateral perturbation," in *Proc. ICLR*, 2023.

[8] H. Inaguma, S. Popuri, I. Kulikov, *et al.*, "Unity: Two-pass direct speech-to-speech translation with discrete units," in *Proc. ACL, VOL.1*, 2023, pp. 15 655–15 680.

[9] C. Yu, Y. Chen, Y. Li, M. Kang, S. Xu, and X. Liu, "Cross-language end-to-end speech recognition research based on transfer learning for the low-resource tujia language," *Symmetry*, vol. 11, no. 2, p. 179, 2019.

[10] R. Tong, L. Wang, and B. Ma, "Transfer learning for children's speech recognition," in *Proceedings of 2017 International Conference on Asian Language Processing (ICALP)*, 2017, pp. 36–39.

[11] R. Zhou, T. Koshikawa, A. Ito, T. Nose, and C.-P. Chen, "Multilingual meta-transfer learning for low-resource speech recognition," *IEEE Access*, vol. 12, pp. 158 493–158 504, 2024.

[12] W.-N. Hsu, D. Harwath, and J. Glass, "Transfer learning from audio-visual grounding to speech recognition," in *Proc. Interspeech 2019*, 2019, pp. 3242–3246.

[13] 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.

[14] T. Tu, Y.-J. Chen, C.-c. Yeh, and H.-Y. Lee, "End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning," *arXiv preprint arXiv:1904.06508*, 2019.

[15] Z. Byambadorj, R. Nishimura, A. Ayush, K. Ohta, and N. Kitaoka, "Text-to-speech system for low-resource language using cross-lingual transfer learning and data augmentation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, p. 42, 2021.

[16] Y. Jia, M. T. Ramanovich, Q. Wang, and H. Zen, "Cvss corpus and massively multilingual speech-to-speech translation," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 6691–6703.

[17] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proc. Interspeech 2021*, 2021, pp. 2426–2430.

[18] J. Tiedemann, M. Aulamo, D. Bakshandaeva, *et al.*, "Democratizing neural machine translation with OPUS-MT," *Language Resources and Evaluation*, no. 58, pp. 713–755, 2023, ISSN: 1574-0218. DOI: 10.1007/s10579-023-09704-w.

[19] V. Pratap, A. Tjandra, B. Shi, *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[20] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, and J. Pino, "Fairseq s2t: Fast speech-to-text modeling with fairseq," in *Proceedings of the 2020 Conference of the Asian Chapter of the Association for Computational Linguistics (AACL): System Demonstrations*, 2020.

[21] M. Post, "A call for clarity in reporting bleu scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, 2018, pp. 186–191.