Direct Vs Cascaded Speech-to-Speech Translation Using Transformer

Lalaram Arya(✉), Amartya Roy Chowdhury, and S. R. Mahadeva Prasanna

Indian Institute of Technology Dharwad, Dharwad 580011, India
{202021004,amartya.chowdhury,prasanna}@iitdh.ac.in

Abstract. Direct speech-to-speech translation (DS2ST) is a process of translating speech from one language to another without using a written form of the language. Most of the works attempted for DS2ST utilized the auxiliary network and knowledge from the written form of the language directly or indirectly to improve the performance. This work proposes a transformer-based sequence-to-sequence model to perform the DS2ST task without an auxiliary network. Also, a comparative study is made with a cascaded system. The experiments are performed with the Prabhupadavani dataset in two languages (Hindi and English). The result shows that with our proposed DS2ST model, a BLEU score of 16.46 is achieved without using any auxiliary information. We also augmented the data with speed perturbation and improved the DS2ST performance BLEU score to 18.58.

Keywords: Direct speech-to-speech translation (DS2ST) · Transformer network · Speech-to-speech translation (S2ST) · Data augmentation

1 Introduction

Speech-to-speech translation (S2ST) is the process of translating the speech of one to another spoken language. More than 40% of the languages of the world do not have a written form of the language [1]. Developing translation technology for such languages is a challenging task. The S2ST system for the languages which have resources (E.g., text and audio dataset) is already explored in detail with the help of the cascaded approach [3]. The traditional way to develop the S2ST system is a cascaded approach. A cascaded S2ST system consists of three modules: Automatic speech translation (ASR), Machine translation (MT), and Text to speech synthesis (TTS) [24]. ASR transforms the source language's speech into text. MT translates source language text (generated via ASR) into target-language text. Finally, the target language text is transformed into the speech of the target language using TTS. This approach has a few issues due to cascading error propagation from one module to another. This technique utilizes the text of both source and target languages to develop the S2ST system. Therefore, building the S2ST system for languages without a written form (spoken-only languages) is challenging [1].

A. Karpov et al. (Eds.): SPECOM 2023, LNAI 14339, pp. 258–270, 2023.
https://doi.org/10.1007/978-3-031-48312-7_21

Recently, researchers have started working to develop a direct speech-to-speech translation (DS2ST) system for spoken-only languages [7,27]. Some attempts have been made, such as implementing a sequence-to-sequence model using a Long short-term memory (LSTM) network called Translatotron [4,11]. In the case of a true DS2ST system, the performance of the system was poor, but with the help of an auxiliary network, it was better, where the written form of the language information was utilized to improve the performance. The extension of the same work was done as Translatotron2 to address the over-generation issues by conditioning the spectrogram synthesizer directly on the output from the auxiliary target phoneme decoder [10]. With the auxiliary network, Taranslatotron2 improves the performance but would still require the target language's text. In these two works, a direct mapping is exploited between the source and the target languages. Some more works were attempted by exploiting the mapping function between the discrete representations between the source and the target language [16,18].

To develop speech technology, the availability of suitable datasets in sufficient quantity is essential. Over the period of time, for the cascaded S2ST approaches, sufficient data has been developed [3]. When it comes to the DS2ST system, the nonavailability of data is severe and lagging significantly as it needs a parallel speech dataset. The works attempted in the field of DS2ST, the synthetic data generated using the TTS system is mostly used [12,14]. To overcome this data scarcity, data augmentation could be a valid solution [5,22]. The augmentation techniques can improve the speaker, gender, channel, and session variability in the data artificially. This incorporates more diverse learning for deep learning networks and improves the system's performance by handling the acoustic variation properly encountered during speech translation. Data augmentation offers a practical and effective solution to handle data scarcity and reduces data collection efforts.

We can learn from the literature that most of the DS2ST systems attempted without a written form of the language lag in performance compared to the cascaded approach [11]. So, a comparative study is needed between the cascaded and the DS2ST systems to understand both systems' best use cases and their pros and cons. In this paper, we are exploring the transformer-based DS2ST system. The 80-dimensional Mel filter bank features are extracted from the raw speech and fed to the transformer-based encoder. The features of target spoken speech are fed to the decoder to train the DS2ST system without using the language's written form, as shown in Fig. 1. Apart from the end-to-end DS2ST training, we also experimented and analyzed the performance with one of the data augmentation techniques, speech perturbation, to resolve data scarcity. We augmented the training and the validation dataset with different perturbation factors α. Augmentation increased the dataset by four times compared to the original dataset. We also developed the cascaded S2ST system to compare its performance with the DS2ST system. The comparative analysis helps us to develop a more accurate and efficient DS2ST system. The experiments show that the performance of the DS2ST system improved significantly compared to the previously attempted

systems, even without the use of the written form of the language. But compared to the MT and cascaded system, it is still lagging. It gives us proof of concept that with the help of advanced deep learning models and more speech data, the task of the DS2ST may be achievable.

The rest of the paper is organised as follows: Sect. 2 describes the proposed transformer-based DS2ST system, cascaded system and data augmentation using speed perturbation. Section 3 details the experimental setup, results, and analysis while developing the systems. Finally, in Sect. 4 we conclude the work and discuss the future direction.

2 Speech-to-Speech Translation System (S2ST)

Traditionally, the S2ST system can be implemented using the cascaded approach where three modules, ASR, MT and TTS, concatenate together. Another one is DS2ST approach, where the spoken speech of one language is translated directly into another without the use of the written form of the language. In this section, we are discussing these two approaches in detail.

2.1 Proposed Direct Speech-to-Speech Translation (DS2ST) Model

Recently, end-to-end speech translation got the attention to translate speech of one language to another directly without using intermediate text. Attempted works approached this problem in two ways. Firstly, by exploiting the direct mapping function that exists between the speech of the source and the target spoken language. Secondly, by exploiting the mapping function that exists between the discrete representation of the source and the target language speech. In this work, a transformed-based sequence-to-sequence modeling has been utilized instead of the LSTM architecture in [11] to speed up the model training and performance.

Transformer Based Speech-to-Speech Translation. Similar to $Speech - Transformer$ for ASR, we developed a multi-head attention-based transformer architecture [8] for the DS2ST task. The model can be formulated as a sequence-to-sequence model [26]. A sequence-to-sequence model converts an input sequence, $x_i = \{x_1, x_2, ...x_t\}$ into an output sequence, $y_i = \{y_1, y_2, ...y_T\}$ where in most of the cases, the sequences are different in lengths ($t \neq T$). In the speech translation task, the input speech of the source language is translated into the target language speech based on conditional probability. 80-dimensional filter bank features are extracted from both the source as well as target speech to feed into the transformer network. The encoder takes filter bank features of the source language and the decoder takes features of the source language as input. Finally, during inference, the speech in the target language is generated, as shown in Fig. 1. A sequence-to-sequence model mathematically can be formulated as follows:

$$e_t = encoder(x_i, h_{t-1}) \tag{1}$$

$$c_t = attention(h, s_{t-1}) \tag{2}$$

$$d_t = decoder(y_{t-1}, s_{t-1}, c_t) \tag{3}$$

where, e_t and d_t are the output from the encoder and decoder, c_t is the context vector calculated by the attention mechanism.

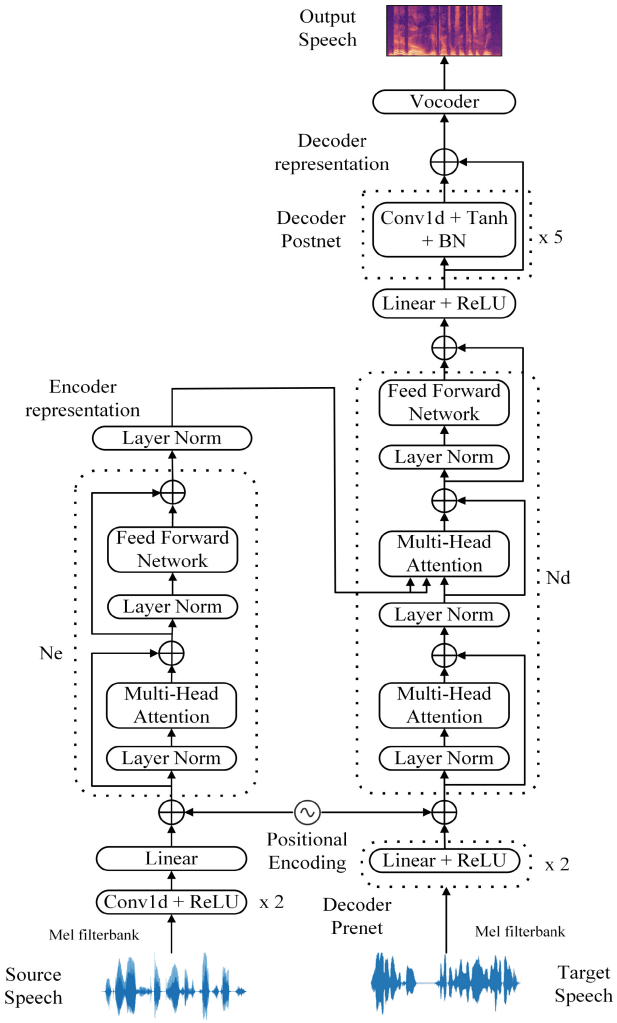

Fig. 1. Transformer based speech-to-speech translation.

Encoder. The encoder takes the input sequence and creates a rich input representation, such as embeddings. In this work, the 80-dimensional filter bank features are extracted as input to the model. The input features are down-sampled

to a quarter of its size by two 1D-CNN layers [15]. Down-sampling makes it easier for the encoder layers to process the intact information with features and reduces memory consumption. The encoder consists of 12 transformer layers with 256− dimensional hidden units. Each multi-head attention block contains eight heads in encoder layers. We have used 1024 dimensional inner states for the feed-forward block followed by Layer-Norm.

Decoder. The decoder takes the output of the encoder (contextual information) and generates an output sequence. In this work, the decoder of our model is inspired by transformer TTS [17], which includes a pre-net and a post-net module. The pre-net consists of two fully connected linear layers. The filter bank features of the target speech are passed into the pre-net as input. The dimension of the pre-net module is set to 256, followed by two ReLU activations. We have also experimented by increasing the dimension to 512. However, this increases the training time instead of making any significant improvement. The decoder consists of 6 layers, keeping the same hyper-parameters as the encoder. To reconstruct the target spectrogram from the decoder, we use a 5−layer 1D-CNN, similar to [23], also called the post-net. The decoder representation from the post-net module is then fed to a vocoder, which converts it into target language speech.

2.2 Cascaded Speech-to-Speech Translation

A cascaded S2ST is a system that translates spoken words from one language into another with the help of the written form of the language. The process of S2ST typically involves three modules: ASR, MT and TTS (Fig. 2).

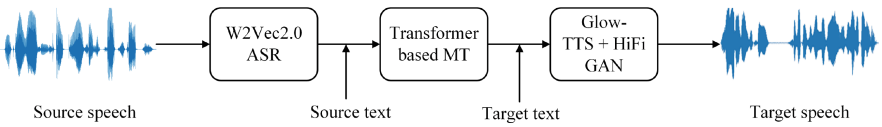

Fig. 2. Cascaded framework of the Speech-to-Speech translation system.

Automatic Speech Recognition (ASR). ASR system recognizes and transcribes the source language speech into the source language text [3]. Wav2Vec 2.0 is a state-of-the-art technique for ASR [6]. By leveraging a self-supervised approach, Wav2Vec 2.0 can effectively learn and extract meaningful speech representations without relying on explicit linguistic supervision. This allows the model to comprehend and identify basic speech elements or phonemes proficiently, even when provided with unlabeled data. We have utilized the pre-trained Wav2Vec 2.0 model and fine-tuned the network weights concerning our dataset. To fine-tune the ASR model, we have used the pertained Vakyansh open source model (CSRIL-23), which is trained on around 10,000 hours of speech collected in 23 Indict languages. The Prabhupadavani dataset consists of 70.5 hours of labeled

speech data and is used for fine-tuning. A fully connected layer is added to the top of the deep network to adapt the pre-trained model for the ASR task. This additional layer was optimized using the connectionist temporal classification (CTC) loss function, a commonly used loss function for sequence labeling tasks in speech recognition. After fine-tuning the model, it achieves a word error rate (WER) of 6.73 in the case of Hindi and 5.91 in the case of the English test dataset.

Machine Translation (MT). MT translates the source language's text into the target language's text [25]. An encoder-decoder transformer-based deep learning network is employed for the MT task in this S2ST system. The transformer network utilizes self-attention and fully connected layers in the encoder and decoder components. The encoder consists of twelve layers, incorporating a multi-head self-attention module and a feed-forward network. Similarly, the decoder consists of six layers, each comprising masked multi-head attention, multi-head attention, and a feed-forward neural network. Transfer learning is employed to enhance the model's performance and adaptability. The training is conducted with the help of Adam Optimizer. Each training batch had a maximum token size of 4096, allowing for efficient computational processing. Byte-pair encoding (BPE) tokenization was applied to all the models, which helps to create smaller dictionaries and enhances vocabulary handling capabilities. By utilizing the transformer network architecture and BPE tokenization, the MT module of the speech-to-speech translation system effectively translates the transcribed text generated by the ASR module into the target language text.

Text-to-Speech Synthesis (TTS). The TTS system synthesizes the speech in the target language from the text translated by the MT in the cascaded S2ST system. To perform this task, the Glow-TTS has been adapted in this work, which was initially proposed for image generation [12]. Using the flow-based generative model, Glow-TTS generates the spectrogram (an intermediate representation of the speech) for the given text. The model uses the invertible convolution and affine coupling layers to learn the conditional distribution of the mel-spectrogram. Also, it leverages the monotonic alignment search (MAS) algorithm to ensure proper and effective alignment during training between the speech and text. After the generation of the spectrogram from the text, a HiFI-GAN vocoder is employed to synthesize it into the target language speech [12]. The vocoder incorporates a generative adversarial network (GAN) that generates high-quality speech. GAN consists of a discriminator and generator, trained using an adversarial learning framework. The model uses a multi-resolution structure, a high-resolution fine-grained feature generator, and a multi-scale discriminator to improve the quality of synthesized speech significantly. This ensures the retention of the acoustic and linguistic characteristics of the desired synthesized speech.

2.3 Data Augmentation

Data augmentation is a technique commonly used to increase the size and diversity of a training dataset artificially. In this work, we have augmented the Prabhupadavani dataset using speed perturbation and investigated the improvement in the performance of DS2ST.

Speed Perturbation. The speed perturbation modified the original speech by resampling the audio signal in the time domain to change its speech rate and duration [9]. If we denote the original audio signal as $x(t)$ and the perturbation factor as α, the resampled audio signal, $y(t)$ can be obtained as follows:

$$y(t) = x(\alpha t) \tag{4}$$

This is equivalent to the following change in the frequency domain:

$$X(f) = \frac{1}{\alpha} x(\frac{1}{\alpha} f) \tag{5}$$

where $X(f)$ and $\frac{1}{\alpha} x(\frac{1}{\alpha} f)$ stand for the respective Fourier transforms of $x(t)$ and $y(t)$. In this way, modifications in speed lead to alterations in the spectral envelope and audio duration.

3 Experiments and Results

In this section, we will discuss our experiments and analyze the results. Firstly, we will discuss the dataset used throughout the experiments and then, with the help of the evaluation matrices, we will analyze the results.

3.1 Data Description

Prabhupadavani dataset is the output of the multi-lingual subtitle generation project of Vanimedia [2]. Under this project, 1080 audio mini-clips were translated, containing conversion, lectures, debates and interviews of swami Prabhupada (founder of an international community called the Society of Krishna Consciousness (ISKCON)). The audio clips contain the conversation about Bhagavad Gita. The primary objective of the project is to translate the audio into 108 languages. Seven hundred dedicated people are working to achieve it. They have to write the translated text according to the given audio clips and the corresponding manually aligned English transcription. The dataset is manually created, so the quality of the corresponding text is excellent. The current release of the dataset consists of 26 languages. In this work, we have taken two languages, Hindi and English, to perform our experiments [21]. For the English language in the dataset, both audio and text are available, but only text is available for Hindi. Table 1 illustrates a few examples of English transcription (source) and the corresponding translations in Hindi from the Prabhupadavani dataset.

Table 1. Some example sentences from Prabhupadavanis dataset

Lang / S.N.	English	Hindi
1	But Caitanya Mahāprabhu said, "No. Anyone can become spiritual master, provided he's conversant with the science.	लेकिन चैतन्य महाप्रभु नें कहा, "नहीं । कोई भी आध्यात्मिक गुरु बन सकता है, अगर वह विज्ञान के साथ परिचित है ।
2	After all, Kṛṣṇa is giving maintenance to everyone.	हिंदी भाषा का उपयोग करके आप अपने दस्तावेज़ में हिंदी पाठ लिख सकते हैं।
3	Prabhupāda: Māyā is nothing. It is a forgetfulness. That's all.	प्रभुपाद: माया कुछ भी नहीं है । यह एक विस्मरण है । बस ।।
4	Punaḥ punaś carvita-carvaṇānām (SB 7.5.30). Adānta-gobhir viśatā tamisra	पुन: पुनश चर्वित-चर्वणानाम (श्रीमद भागवतम ७.५.३०) । अदान्त-गोभिर वि-शताम तमिश्रम

To perform the S2ST task, we have utilized the fine-tuned TTS described in Sect. 2.2 to generate the Hindi speech from the available transcriptions. We have also observed that the quality of the available source speech with the data in English is noisy as it is recorded in a very open environment during conversation, discussion and lectures. Between the conversion, multiple speaker's voices, long pauses, noise and many filler words are present. The mapping of such speech with the TTS-generated speech is complicated. To overcome such issues during DS2ST, we also generated the source English speech using TTS.

The dataset contains around 53, 398 utterances for each language. The train set consists of 51, 301 utterances, while the test and dev sets contain 1, 048 and 1, 049 utterances, respectively. For further insights, see Table 2.

Table 2. Statistics of Prabhupadavani dataset in terms of number of sentences.

Language	Train	Dev	Test	Total
English	51301	1048	1049	53398
Hindi	51301	1048	1049	53398

3.2 Evaluation Metrics

The evaluation of our system involves assessing both the translation quality and the speech quality of the generated output. We follow the setup outlined in [11] to evaluate the translation quality by applying ASR to the speech output. We compute BLEU scores by comparing the ASR-decoded text to the reference

translations. We utilize an open-sourced Hndi ASR model for the ASR component that combines Wav2Vec 2.0 pre-training with self-supervised techniques [6]. This model is pre-trained on 23 Indian languages and fine-tuned on the Prabhupadavani dataset. It achieves a Word Error Rate (WER) of 6.73 on the TTS-generated test set. We normalize the reference text to ensure consistency before computing the BLEU scores using SACREBLEU [20].

3.3 Experimental Set-Up

We have trained the proposed model on the Prabhupadvani corpus [21]. The 80-dimensional Mel-filterbank is computed with $25ms$ of the window and $10ms$ of overlap duration for the input speech. We have also applied spec augment [19]. The down-sampling task consists of two Conv1D layers with a kernel size of 5, stride of 2 and 1024 channels. The encoder and decoder contain eight attention layers with an embedding dimension of 256. We have trained the model for 3000 epochs using Adam optimizer [13] with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. We have also applied a label smoothing of 0.2 and a learning rate of 10^{-5}. We have used an inverse square root learning rate scheduler with $10k$ warm-up steps. The attention dropout and the feed-forward dropout are kept constant at 0.1 throughout the network, while a dropout of 0.5 is applied in the decoder post-net module.

3.4 Results and Discussion

We evaluated the performance of DS2ST and cascaded translation systems developed with Prabhupadavani datasets, which contain speech generated by the TTS system in a male voice. We employed the SACREBLEU metric on ASR transcriptions derived from the translated speech to assess the translation performance. The text obtained from ASR transcriptions was converted to lowercase, excluding punctuation marks and apostrophes. It's important to note that the BLEU score is traditionally used for evaluating machine translation systems based on reference translations. However, in this case, we used ASR as an intermediary step to convert the translation speech to text, which introduces potential errors. Consequently, the BLEU scores obtained from ASR transcriptions represent a distorted estimate, providing a lower bound on the translation quality. To ensure a fair comparison, we utilized an ASR model from Ekstep that was pre-trained on the 23 languages and fine-tuned with the Prabhupada dataset. This ASR model served as a common evaluation framework for all the models, including the baseline models. Furthermore, to generate the audio waveforms from the predicted mel-spectrograms, we employed the HiFi-GAN vocoder consistently across all models. Overall, this evaluation methodology allowed us to objectively compare the translation performance of different speech-to-speech translation models using a standardized ASR model and estimate their translation quality.

Objective Evaluation. In this work, we have performed four experiments. Firstly, we have proposed a DS2ST system, and the system's performance was analyzed on Prabhupadavani's unseen test dataset. We can see from the Table 3 that the system can translate English speech into Hindi without knowing the language's written form. The BLEU score we have achieved is 16.46 without any auxiliary decoder, which was 0.4 in the case of the Translatotron model.

Table 3. Performance analysis of the DS2ST vs cascaded system.

Model	BLEU
MT	37.31
Cascaded model	30.90
DS2ST without Aug	16.46
DS2ST with Aug	18.58

Secondly, we developed a cascaded system to compare the performance of our proposed DS2ST system. From the Table 3, we see that in comparison to the DS2ST system, cascaded leads with a high margin of BLEU score of 14.44. The difference in the BLEU score indicates the difficulty in exploiting the mapping function without the written form of the language. Thirdly, we can see a performance gap when we analyze the result of MT compared to the cascaded system from the Table 3. This again indicates that when we are embedding the ASR, generating the source and the target text degrades the performance by a 6.41 BLEU score. Finally, we have checked the data augmentation technique's capability in the DS2ST system. In this experiment, we have augmented the data by four times with speed perturbation. Table 3 indicates the significant improvement in performance by BLEU score 2.12.

Subjective Evaluation. In the subjective study, we conducted an intelligibility test to assess the performance of different translation models. We randomly selected 20 translated utterances from each translation model under consideration to ensure an unbiased evaluation. During the intelligibility test, the selected utterances from different models were played randomly to the listeners. Each listener was then asked to listen to each utterance once and write down the sentence they heard based on their understanding.

To evaluate the performance, we compared the sentences written by each listener with the ground truth, which consists of the correct original text sentences. The percentage accuracy was calculated by determining the number of words in the sentences that were accurately transcribed out of the total number of words in the sentences played to the listeners. This process helps us to gauge the intelligibility of the translations effectively and provides valuable insights into the effectiveness of the various translation models under study.

Table 4. Performance analysis with intelligibility score (IS) of the DS2ST vs cascaded system.

Model	%IS
Cascaded model	32.79
DS2ST without Aug	17.35
DS2ST with Aug	18.99

The findings from Table 4 indicate that human performance in the intelligibility test aligns with the trends observed in the objective study. Furthermore, it is noticeable that there is a slight discrepancy in the scores between the human evaluation and the BLEU metric. This difference may be attributed to higher human intelligibility and more comprehensive than the assessment provided by automated metrics like BLEU.

4 Conclusion and Future Direction

In this work, we have developed a transformer-based DS2ST system. Experiments are performed with the Prabhupadavani dataset, where we considered English as the source language and Hindi as the target language. We also developed a cascaded S2ST system using the same dataset to compare the performance. A comparative study is made between the DS2ST and the cascaded system. The results show that the system's performance has improved significantly by the BLEU score of 16.60 compared to the earlier DS2ST system (Translatotron 0.4). At the same time, when we compare the performance between the DS2ST and the cascaded system, the DS2ST system still lags significantly. Also, by comparing the cascaded system with MT, we can see that cascading the ASR degrades the system performance by a BLEU score of 6.60. From this, we can conclude that, as of now, the performance of the DS2ST system is lagging compared to the cascaded system, but it is possible to develop the DS2ST system without the use of the written form of language with the advancement in deep learning networks. In the future, we would like to explore other deep learning models and also more data to further improve the performance of the DS2ST system.

Acknowledgment. We extend our gratitude to the "AnantGanak" high-performance computation (HPC) facility at IIT Dharwad for facilitating our research experiments.

References

1. Languages of the world. https://www.ethnologue.com. Accessed 18 Sep 2023
2. Vanimedia. https://vanimedia.org/wiki. Accessed 21 Sep 2023
3. Alhari, S., et al.: Automatic speech recognition: systematic literature review. IEEE Access **9**, 131858–131876 (2021)

4. Arya, L., Agarwal, A., Mishra, J., Prasanna, S.R.M.: Analysis of layer-wise training in direct speech to speech translation using BI-LSTM. In: 2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), pp. 1–6 (2022)
5. Arya, L., Agarwal, A., Prasanna, S.R.M.: Investigation of data augmentation techniques for BI-LSTM based direct speech to speech translation. In: 2023 National Conference on Communications (NCC), pp. 1–6 (2020)
6. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: a framework for self-supervised learning of speech representations. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 12449–12460. Curran Associates, Inc. (2020)
7. Chen, P.J., et al.: Speech-to-speech translation for a real-world unwritten language. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 4969–4983. Association for Computational Linguistics (2023)
8. Dong, L., Xu, S., Xu, B.: Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5884–5888 (2018)
9. Geng, M., et al.: Investigation of data augmentation techniques for disordered speech recognition. In: Interspeech 2020. ISCA (2020)
10. Jia, Y., Ramanovich, M.T., Remez, T., Pomerantz, R.: Translatotron 2: high-quality direct speech-to-speech translation with voice preservation. In: Proceedings of the 39th International Conference on Machine Learning. vol. 162, pp. 10120–10134. PMLR (17–23 Jul 2022)
11. Jia, Y., et al.: Direct speech-to-speech translation with a sequence-to-sequence model. In: INTERSPEECH, pp. 1123–1127 (2019)
12. Kim, J., Kim, S., Kong, J., Yoon, S.: Glow-TTS: a generative flow for text-to-speech via monotonic alignment search. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20 (2020)
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings (2015)
14. Kong, J., Kim, J., Bae, J.: HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20 (2020)
15. LeCun, Y., Bengio, Y.: Convolutional networks for images, speech, and time series, pp. 255–258, MIT Press (1998)
16. Lee, A., et al.: Direct speech-to-speech translation with discrete units. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 3327–3339. Association for Computational Linguistics (2022)
17. Li, N., Liu, S., Liu, Y., Zhao, S., Liu, M.: Neural speech synthesis with transformer network. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence. AAAI'19, AAAI Press (2019)
18. Li, X., Jia, Y., Chiu, C.C.: Textless direct speech-to-speech translation with discrete speech representation. In: ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023)
19. Park, D., et al.: Specaugment: a simple data augmentation method for automatic speech recognition. In: Interspeech, pp. 2613–2617. ISCA (09 2019)

20. Post, M.: A call for clarity in reporting BLEU scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers, pp. 186–191. Association for Computational Linguistics, Brussels, Belgium (2018)
21. Sandhan, J., Daksh, A., Paranjay, O.A., Behera, L., Goyal, P.: Prabhupadavani: a code-mixed speech translation data for 25 languages. LaTeCH-CLfL, pp. 24–29 (2022)
22. Shahnawazuddin, S., Adiga, N., Kathania, H.K., Sai, B.T.: Creating speaker independent ASR system through prosody modification based data augmentation. Pattern Recogn. Lett. **131**, 213–218 (2020)
23. Shen, J., et al..: Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech nd Signal Processing (ICASSP), pp. 4779–4783 (2018)
24. Shimizu, T., Ashikari, Y., Sumita, E., Zhang, J., Nakamura, S.: NICT/ATR Chinese-Japanese-English speech-to-speech translation system. Tsinghua Sci. Technol. **13**(4), 540–544 (2008)
25. Tan, Z., et al.: Neural machine translation: a review of methods, resources, and tools. AI Open **1**, 5–21 (2020)
26. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
27. Zhang, C., Tan, X., Ren, Y., Qin, T., jun Zhang, K., Liu, T.Y.: Uwspeech: speech to speech translation for unwritten languages. In: AAAI Conference on Artificial Intelligence, pp. 14319–14327 (2020)