Interspeech 2022
18-22 September 2022, Incheon, Korea

Leveraging Pseudo-labeled Data to Improve Direct Speech-to-Speech Translation

Qianqian Dong[1,*], *Fengpeng Yue*[2,*,†], *Tom Ko*[1], *Mingxuan Wang*[1], *Qibing Bai*[2,†], *Yu Zhang*[2,3]

[1]ByteDance AI Lab
[2]Department of Computer Science and Engineering,
Southern University of Science and Technology, Shenzhen, China
[3]Peng Cheng Laboratory, Shenzhen, China

{dongqianqian,tom.ko,wangmingxuan.89}@bytedance.com,
{11930381,12032871}@mail.sustech.edu.cn, yu.zhang.ust@gmail.com

Abstract

Direct Speech-to-speech translation (S2ST) has drawn more and more attention recently. The task is very challenging due to data scarcity and complex speech-to-speech mapping. In this paper, we report our recent achievements in S2ST. Firstly, we build a S2ST Transformer baseline which outperforms the original Translatotron. Secondly, we utilize the external data by pseudo-labeling and obtain a new state-of-the-art result on the Fisher English-to-Spanish test set. Indeed, we exploit the pseudo data with a combination of popular techniques which are not trivial when applied to S2ST. Moreover, we evaluate our approach on both syntactically similar (Spanish-English) and distant (English-Chinese) language pairs. Our implementation is available at `https://github.com/fengpeng-yue/speech-to-speech-translation`.

Index Terms: speech translation, speech-to-speech translation, pseudo-labeling

1. Introduction

The speech-to-speech translation (S2ST) task translates speech from the source language into speech in the target language. The conventional S2ST is implemented by a cascaded system [1], which includes three components: automatic speech recognition (ASR), text-to-text machine translation (MT), and text-to-speech synthesis (TTS). Like cascaded speech-to-text translation (ST), the two main shortcomings of cascaded S2ST are time delay and error accumulation. The end-to-end (E2E) approach, which jointly optimizes all components with a single model, effectively alleviates these problems. With the success of end-to-end ST, the community starts to move on to direct S2ST, a more difficult task.

Translatotron [2] is one of the pioneering works in direct S2ST, which is conducted by adopting multi-task learning [3]. It verifies that the end-to-end method can obtain reasonable translation quality and generate intelligible speech. After that, [4] proposes Translatotron2 to improve the robustness of the predicted speech and retain the source speaker's voice in the translated speech better. For distant language pairs, [5] explores the multi-step training with Transcoder. The pre-trained MT and TTS encoders are used as the teacher model to facilitate direct S2ST in learning complex linguistic and modal transitions.

On the other hand, [6] aims to build a direct S2ST for unwritten languages. Instead of predicting continuous spectrograms, [6] predicts discrete units learned from self-supervised representations of the target speech. Multi-task learning can be conducted either with text data or not. Furthermore, a textless S2ST system is proposed by [7] and can be trained without any text data. At the same time, it can generate multi-speaker target speech by training on real-world S2ST data.

One of the challenges to train the end-to-end S2ST system is data scarcity, as collecting speech translation data is expensive. To tackle this challenge, [8] proposes an automatic data mining method to perform speech-to-speech mining. However, no work has proven that it is feasible to train the S2ST on these automatically mined data. Meanwhile, the pseudo-labeled data effectively improve the performance of E2E ST [9, 10] when real-world data is limited. However, currently there are few investigations on direct S2ST. This motivates us to investigate leveraging a large amount of pseudo-labeled data to enhance the performance for S2ST when only limited paired data can be obtained. As more and more large-scale ASR resources are open-sourced, in this paper, we utilize well-trained MT and TTS model to convert ASR data to pseudo-labeled S2ST data. To evaluate different methods for using pseudo-labeled data, we conduct experiments on similar language pairs (i.e., Spanish-English) and distant language pairs (i.e., English-Chinese). Our contributions are three-fold:

- We build a strong Transformer-based Translatotron baseline that outperforms the original Translatotron. We report the thorough evaluation of the effects of hyperparameter tuning.
- We examine the effectiveness of using pseudo-labeled data with pre-training and different fine-tuning strategies. By utilizing pseudo-labeled data and the prompt-tuning technique [11], our best model achieves new state-of-the-art results on the Fisher dataset [12].
- Except for Spanish-English, we evaluate our approach on English-Chinese, which is a syntactically distant language pair, to facilitate the S2ST research in different language communities.

The rest of the paper is organized as follows. Section 2 describes our implementation of Transformer-based Translatotron. In Section 3, we introduce our pseudo-labeling approach. Section 4 describes the experiments and presents the analysis. Section 5 concludes our work.

2. Transformer-based Translatotron

As shown in Figure 1, our model structure follows the Translatoron [2]. We adopt the Transformer [13] instead of LSTM as

* Equal contribution.
† Work done during internship at Bytedance.

10.21437/Interspeech.2022-10011

the backbone. We find that Transformer-based Translatotron can get much better performance than original Translatoron with careful parameter tuning.

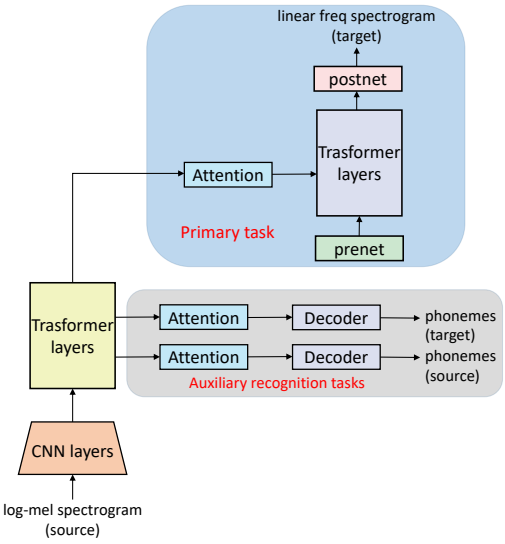

Figure 1: *Transformer-based Translatotron translates speech from the source language (bottom left) into the speech of the target language (top right). The auxiliary tasks help learn the speech-to-speech translation. Compared with the original Translatotron, we focus on generating target speech on a single speaker without using the speaker encoder.*

2.1. Transformer Encoder

Subsampling the speech feature in the encoder is an effective way to help the model pick up attention during training. In our model, the 80-channel mel-spectrogram input features are subsampled to a quarter of the original size by two convolutional layers and then fed into the Transformer layer. The Transformer encoder has 12 Transformer layers with 512-dimension hidden units in each layer and 8 heads in each multi-head attention block. For the position-wise feed-forward networks, we use 2048 dimensional inner states. During training, SpecAugment [14] is applied as the data augmentation strategy.

2.2. Transformer Decoder

Similar to the Transformer TTS [15], the decoder that predicts the spectrogram in the target language includes pre-net, Transformer layers, and post-net components. By following the setting in Translatotron, the dimension of the pre-net bottleneck is set to 32. Based on the performance on the development set, we set the reduction factor [16] to 4 for the output feature frames. The Transformer decoder has 6 Transformer layers with the same hyperparameters as the encoder's Transformer layers.

2.3. Multi-task Learning

Following Translatotorn, we also employ multi-task learning training strategy. In addition to the primary task (i.e., speech-to-speech translation), two additional tasks (i.e., speech recognition and speech-to-text translation) are included in Translatotron. The two auxiliary tasks play an important role in Translatotron, and they are conducted by two additional decoders. The auxiliary decoders take the intermediate hidden representation of the encoder to predict phoneme sequences. Intuitively, the shallow encoder layers represent the source linguistic content, while the deep layers encode more information about the target linguistic content. During training, the weighted auxiliary losses are added to the overall training loss. In Section 4, we demonstrate that the performance of the primary task is sensitive to the hyperparameters of the auxiliary tasks. Table 1 shows the details about the hyperparameter setting of our best practice.

Table 1: *Model hyperparameters for Fisher and TedEn2Zh datasets.*

#Param.	**Fisher**	**TedEn2Zh**
Input / output sample rate (Hz)	8k/24k	16k/24k
Transformer Encoder	12 x 512	12 x 512
Transformer Decoder	6 x 512	6 x 512
Auxiliary Transformer Decoder	1 x 64	4 x 64
source / target encoder layer	6 / 9	4 / 9
source / target loss weight	0.3 / 0.3	0.3 / 0.3
Learning rate	0.006	0.0015
Warmup steps	4000	4000
Dropout	0.1	0.1
Batch tokens	60000	45000

3. Pseudo Translation Labeling

As real S2ST data is very limited nowadays, pseudo-labeling is an intuitive approach to alleviate the problem. Most existing work [2, 4, 5] converts ST data to S2ST data with a TTS system and conducts their experiments. However, real ST data is also limited, and the data scarcity problem remains severe. In this paper, we extend the pseudo-labeling approach by converting the ASR data into ST data with a MT system and then into S2ST data.

In our work, we first create a primary dataset $\mathcal{A}$ from a ST corpus. $\mathcal{A} = \{\mathbf{s_{src}}, \mathbf{t_{src}}, \mathbf{t_{tgt}}, \mathbf{s'_{tgt}}\}$ represents {real source speech, real transcription, real translation, pseudo target speech}. Then we create a secondary dataset $\mathcal{B}$ from an ASR corpus. $\mathcal{B} = \{\mathbf{s_{src}}, \mathbf{t_{src}}, \mathbf{t'_{tgt}}, \mathbf{s'_{tgt}}\}$ represents {real source speech, real transcription, pseudo translation, pseudo target speech}. Here, dataset $\mathcal{B}$ is much larger than $\mathcal{A}$ in scale. We examine the effectiveness of using the secondary dataset with pre-training and different fine-tuning strategies. As the target translation sequences of dataset $\mathcal{B}$ are generated by MT, we name our approach as pseudo translation labeling (PTL).

3.1. Pre-training and Fine-tuning

We first use dataset $\mathcal{B}$ to pre-train the encoder so that the model can learn a better representation. The pre-training is done with an equal weight for two tasks: ASR and ST, which make use of $\{\mathbf{s_{src}}, \mathbf{t_{src}}\}$ and $\{\mathbf{s_{src}}, \mathbf{t'_{tgt}}\}$, respectively. The two tasks share representations from the same encoder but with different decoders. The pre-trained encoder and the decoders are reused in the upcoming fine-tuning stage. Dataset $\mathcal{A}$ is then used to optimize the overall model to carry out the basic fine-tuning.

3.2. Mixed-tuning

In order to further improve the model, we conduct fine-tuning with a mixture of dataset $\mathcal{A}$ and $\mathcal{B}$. As dataset $\mathcal{B}$ is much larger

than $\mathcal{A}$, we duplicate the samples in the primary set to balance the overall distribution. In this work, we assume that the distribution of the target test set is close to the distribution of the primary training set. Thus, we explicitly apply upsampling to prevent the model from getting biased to the secondary training set.

3.3. Prompt-tuning

In order to enhance the ability of the model to learn the difference between various data sources, we adopt the "pre-train, prompt, and predict" [11] paradigm. Based on pre-training, we take the category of the datasets (including "<primary>" and "<secondary>") as a prompt, and attach it to the input features of each sample in the form of the predefined embeddings during the prompt-tuning stage. With the explicit textual prompt, we can manipulate the model's behavior in the inference stage.

4. Experiments

4.1. Datasets

We conduct experiments on two language pairs (i.e., Spanish to English and English to Chinese). The former belongs to the same language family, while the latter belongs to a different language family. We construct S2ST paired data based on the two ST datasets, Fisher Spanish[12] and TedEn2Zh[17], by using the in-house TTS to synthesize the target speech from the translation sequences. We utilize the in-house MT to convert ASR data that includes a subset of Gigaspeech [18] (Giga-Sub) and the Spanish subset of multilingual LibriSpeech [19] (MLS_Es) to pseudo-labeled ST data, and then use the same in-house TTS to synthesize the target speech from the pseudo translation sequences. As described in Section 3, for the Spanish-to-English language pair, Fisher and MLS_Es are the primary and secondary datasets, respectively. TedEn2Zh and Giga-Sub are the the primary and secondary datasets for the English-to-Chinese language pair, respectively. Table 2 shows the statistics of each dataset.

Table 2: *Statistics of training data for Spanish-to-English (Es → En) and English-to-Chinese (En → Zh). "Mixed" includes audiobook, podcast, and YouTube. All duration statistics are based on segmented audios.*

Dataset	Es → En		En → Zh	
	primary	secondary	primary	secondary
Data source	Fisher	MLS_Es	TedEn2Zh	Giga-Sub
Source hour	171	918	524	1566
Target hour	146	738	467	1478
Utterance	130K	220K	294K	1.1M
Sampling rate	8 kHz	16 kHz	16 kHz	16 kHz
Domain	Conversation	Audiobook	Lecture	Mixed

4.2. Implementation Details

Data Preprocessing To be consistent with the sampling rate (SR) of Fisher, we downsample MLS_Es to 8 kHz. Our acoustic features are 80-channel mel-spectrogram extracted from source speech as input and 80-channel mel-spectrogram extracted from target speech as output. We use phonemes from the source and target languages as modeling units for the two auxiliary tasks. We filter out the code-switched texts and those texts for which TTS fails to generate speech.

Evaluation Following previous work, we use the BLEU score as an objective evaluation metric to measure the translation accuracy and the mean opinion score (MOS) as a subjective evaluation metric to measure the naturalness of pronunciation for predicted target speech. We utilize the pre-trained ASR to recognize the predicted target speech and then calculate the BLEU score between the resulting transcripts and ground-truth reference translations. For English ASR, we use the `Wav2vec2`[20] and `CTC` model[1] from `Huggingface`[21], which is pretrained and fine-tuned on Libri-Light[22] and 960 hours of Librispeech[23] corpus. For Chinese ASR, we use the attention-based `Conformer`[24] model[2] from `Wenet`[25] trained on the AISHELL[26] corpus. We report case-insensitive detokenized BLEU scores calculated by `sacrebleu`[3] on Fisher dataset. Meanwhile, we report character-level BLEU scores on the TedEn2Zh dataset.

Cascade S2ST Our cascade S2ST is built by cascading E2E ST and TTS. We train a Transformer-based E2E ST following the model setting in [27] with ASR pre-training. In the cascade S2ST system, we take the predicted translation of ST to generate the target speech using the in-house TTS.

4.3. Main Results

In this section, we illustrate the effectiveness of our method from the perspective of objective evaluation. Translatotron-T represents the Transformer-based Translatotron in all tables.

4.3.1. Objective Evaluation

Table 3: *Performance on the dev and test sets of Fisher Spanish-English dataset.*

Method	dev-BLEU	dev2-BLEU	test-BLEU
Translatotron [28]	24.8	26.5	25.6
+ Encoder PT	30.1	31.5	31.1
Translatotron-2 [4]	-	-	37.0
+ ConcatAug	-	-	40.3
UWSpeech [29]	-	-	9.4
S2UT [6]	-	-	39.9
Translatotron-T	32.1	32.8	32.0
+ PTL	**42.4**	**43.3**	**43.6**
Cascade S2ST [28]	39.4	41.2	41.4
Cascade S2ST [4]	-	-	43.3
Cascade S2ST	44.3	45.4	45.1
Ground truth [28]	82.8	83.8	85.3
Ground truth [4]	-	-	88.6
Ground truth	88.1	88.6	89.8

Performance on Fisher dataset We present our results in Table 3. It demonstrates that our baseline results surpass the original Translatotron results on Fisher. Moreover, we adopt the PTL described in Section 3 on our baseline, which improves the BLEU score significantly. Overall, our method can achieve an

[1] `https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self`

[2] `https://github.com/wenet-e2e/wenet/tree/main/examples/aishell`

[3] `https://github.com/mjpost/sacrebleu`

improvement of nearly 3 BLEU scores compared with the previous best end-to-end model, Translatotron-2 with "ConcatAug" data augmentation.

In Table 4, we report the performance on auxiliary tasks under different methods. The results show that the auxiliary tasks and the primary task have a positive correlation in performance. This indicates that using external data to improve the performance of auxiliary tasks will also benefit the S2ST, which is consistent with the motivation of pre-training auxiliary tasks.

Table 4: *Performance of auxiliary tasks on the test set of Fisher Spanish-English dataset. "S-PER" means phoneme error rate (PER) of auxiliary ASR task on test set. "Tp-BLEU" means phone-based BLEU of auxiliary ST task on test set.*

Method	S-PER(↓)	Tp-BLEU	test-BLEU
Translatotron-T	16.10	55.68	32.0
+ PTL	**13.95**	**61.96**	**43.6**

Performance on TedEn2Zh dataset English-Chinese translation is a more difficult task because the two languages are much different in grammar and syntax. As shown in Table 5, we report the results of Transform-based Translatotron on the Ted2Zh dataset. The BLEU score of the baseline is merely 11.2 because the predicted results contain a large amount of unintelligent speech. The PTL can bring about 7 points of BLEU improvement, which significantly enhances translation accuracy. Our best model even surpasses the cascade system.

Table 5: *Performance on the dev and test set of TedEn2Zh dataset. "S-PER" means PER of auxiliary ASR task on test set. "Tp-BLEU" means phoneme-based BLEU of auxiliary ST task on test set.*

Method	S-PER(↓)	Tp-BLEU	dev-BLEU	test-BLEU
Translatotron-T	13.06	44.57	7.0	11.2
+ PTL	**11.26**	**50.73**	**12.2**	**20.8**
Cascade S2ST	-	-	11.8	19.7
Ground truth	-	-	82.9	94.9

Parameters for Auxiliary Task We explore the influence of the hyperparameters of auxiliary tasks on the performance of S2ST on the Fisher Spanish dataset, and the results are shown in Table 6. We find that the performance of the model with the size of the smaller auxiliary decoder is better. We conjecture that it will force the encoder shared by the auxiliary task and the primary task to learn more helpful information, which is more conducive to the training of the primary task.

Effect of different methods As shown in Table 7, we compare different methods of the pseudo translation labeling approach. When we apply the pre-training to the auxiliary tasks by pseudo-labeled data (Method-I), the BLUE scores significantly improve compared with the baseline. Further, based on pre-training, mixed-tuning (Method-II) improves 2.8 BELU on Fisher and 5.5 BELU on TedEn2Zh. As shown in Table 2, there is an obvious mismatch between the primary data and the secondary data in the two language pairs. The prompt-tuning (Method-III) helps the model distinguish different data sources, and further gains can be obtained on both language pairs.

Table 6: *Effect of the parameters for the auxiliary tasks (ASR, ST) on the test set of Fisher dataset. "#Pos" represents the source and target encoder layer. "#Num" and "#Dim" mean the number and dimension of the Transformer layers of two auxiliary decoders , respectively.*

Exp.	#Pos	#Num	#Dim	test-BLEU
I	(4,9)	(4,4)	(128,128)	24.2
II	(4,9)	(3,3)	(128,128)	25.7
III	(6,9)	(3,3)	(128,128)	28.4
IV	(6,9)	(1,1)	(64,64)	**32.0**

Table 7: *Comparison on the effectiveness of the three methods. The BLEU scores are reported on the test sets of Fisher and TedEn2Zh datasets. "Pre-training" is conducted on the dataset $\mathcal{B}$ for all three methods.*

Model	Pre-training	Fine-tuning	Fisher	TedEn2Zh
Translatotron-T	-	-	32.0	11.2
Method-I	✓	✓	39.2	14.8
Method-II	✓	Mixed-tuning	43.0	20.3
Method-III	✓	Prompt-tuning	**43.6**	**20.8**

4.3.2. Subjective Evaluation

In Table 8, we use the Hifi-GAN [30] vocoder to synthesize audios from predicted spectrograms and conduct the MOS test to evaluate the naturalness of audios. The gain from our PTL approach on MOS is consistent with BLEU. In particular, our method significantly enhances the intelligibility of audios on the TedEn2Zh dataset (English-Chinese).

Table 8: *Naturalness MOS Evaluation on test sets of the two datasets. The ground truth for both datasets are synthetic target speech from the in-house TTS.*

Method	Fisher	TedEn2Zh
Translatotron-T	3.40±0.16	2.59±0.18
+ PTL	**3.59**±0.17	**3.20**±0.18
Ground truth	3.87±0.19	3.95±0.18

5. Conclusions

In this paper, we first build a strong Transformer-based Translatotron baseline for direct S2ST, which obviously outperforms the original Translatotron. As the S2ST performance is sensitive to the hyperparameters of the auxiliary decoder, we have made careful tuning to get the best performance. Furthermore, to tackle data scarcity problem, we examine the effectiveness of employing pseudo-labeling with pre-training and various fine-tuning strategies to utilize the unlabeled data. Our system can achieve new state-of-the-art performance on Fisher (Spanish-English) dataset. Finally, we report the performance of TedEn2Zh (English-Chinese) dataset to facilitate the direct S2ST research on more language pairs.

6. References

[1] A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld, and P. Zhan, "Janus-iii: Speech-to-speech translation in multiple languages," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1997, pp. 99–102.

[2] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model," in *Proc. Interspeech 2019*, 2019, pp. 1123–1127.

[3] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[4] Y. Jia, M. T. Ramanovich, T. Remez, and R. Pomerantz, "Translatotron 2: Robust direct speech-to-speech translation," *arXiv preprint arXiv:2107.08661*, 2021.

[5] T. Kano, S. Sakti, and S. Nakamura, "Transformer-based direct speech-to-speech translation with transcoder," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 958–965.

[6] A. Lee, P.-J. Chen, C. Wang, J. Gu, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang, J. Pino *et al.*, "Direct speech-to-speech translation with discrete units," *arXiv preprint arXiv:2107.05604*, 2021.

[7] A. Lee, H. Gong, P.-A. Duquenne, H. Schwenk, P.-J. Chen, C. Wang, S. Popuri, J. Pino, J. Gu, and W.-N. Hsu, "Textless speech-to-speech translation on real data," *arXiv preprint arXiv:2112.08352*, 2021.

[8] P.-A. Duquenne, H. Gong, and H. Schwenk, "Multimodal and multilingual embeddings for large-scale speech mining," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[9] Y. Jia, M. Johnson, W. Macherey, R. J. Weiss, Y. Cao, C.-C. Chiu, N. Ari, S. Laurenzo, and Y. Wu, "Leveraging weakly supervised data to improve end-to-end speech-to-text translation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7180–7184.

[10] A. D. McCarthy, L. Puzon, and J. Pino, "Skinaugment: auto-encoding speaker conversions for automatic speech translation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7924–7928.

[11] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *arXiv preprint arXiv:2107.13586*, 2021.

[12] M. Post, G. Kumar, A. Lopez, D. Karakos, C. Callison-Burch, and S. Khudanpur, "Improved speech-to-text translation with the fisher and callhome spanish-english speech translation corpus," in *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*, 2013.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[14] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019.

[15] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.

[16] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *Proc. Interspeech 2017*, pp. 4006–4010, 2017.

[17] Y. Liu, H. Xiong, J. Zhang, Z. He, H. Wu, H. Wang, and C. Zong, "End-to-end speech translation with knowledge distillation," *Proc. Interspeech 2019*, pp. 1128–1132, 2019.

[18] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," in *Proc. Interspeech 2021*, 2021.

[19] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," *Proc. Interspeech 2020*, pp. 2757–2761, 2020.

[20] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[21] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[22] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7669–7673.

[23] 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.

[24] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *Proc. Interspeech 2020*, pp. 5036–5040, 2020.

[25] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei, "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," in *Proc. Interspeech*. Brno, Czech Republic: IEEE, 2021.

[26] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[27] H. Inaguma, S. Kiyono, K. Duh, S. Karita, N. Yalta, T. Hayashi, and S. Watanabe, "Espnet-st: All-in-one speech translation toolkit," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020, pp. 302–311.

[28] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," *Proc. Interspeech 2019*, pp. 1123–1127, 2019.

[29] 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, no. 16, 2021, pp. 14 319–14 327.

[30] 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.