JOINT PRE-TRAINING WITH SPEECH AND BILINGUAL TEXT FOR DIRECT SPEECH TO SPEECH TRANSLATION

Kun Wei[1,†]*, Long Zhou*[2]*, Ziqiang Zhang*[2]*, Liping Chen*[2]*, Shujie Liu*[2]*, Lei He*[2]*, Jinyu Li*[2]*, Furu Wei*[2]

[1]Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University, Xian, China
[2]Microsoft Corporation

ABSTRACT

Direct speech-to-speech translation (S2ST) is an attractive research topic with many advantages compared to cascaded S2ST. However, direct S2ST suffers from the data scarcity problem because the corpora from the speech of the source language to the speech of the target language are very rare. To address this issue, we propose in this paper a Speech2S model, which is jointly pre-trained with unpaired speech and bilingual text data for direct speech-to-speech translation tasks. By effectively leveraging the paired text data, Speech2S is capable of modeling the cross-lingual speech conversion from source to target language. We verify the performance of the proposed Speech2S on Europarl-ST and VoxPopuli datasets. Experimental results demonstrate that Speech2S gets an improvement of about 5 BLEU scores compared to encoder-only pre-training models, and achieves a competitive or even better performance than existing state-of-the-art models[1].

Index Terms— Speech to speech translation, joint pre-training, cross-lingual modeling.

1. INTRODUCTION

Direct speech to speech translation (S2ST) has gained more and more attention from research and industry communities in recent years [1–3]. Traditionally, cascaded S2ST consists of automatic speech recognition (ASR), machine translation (MT), and text to speech synthesis (TTS) tasks. Direct S2ST aims at integrating the above three tasks into an end-to-end model, which translates the speech of one language to the speech of another language directly. Compared to cascaded S2ST, direct S2ST has the following advantages: (1) it is able to alleviate the error propagation problem of pipeline systems; (2) it can retain the emotion, pitch, and prosody information of the speaker to the greatest extent; (3) it has faster reasoning speed and takes up fewer storage resources.

However, data scarcity is the biggest problem of direct S2ST tasks [4]. At present, there is very little parallel S2ST data though lots of efforts [5–7]. To alleviate this problem, a line of work tries to leverage pseudo data to improve direct S2ST [3, 8]. They usually convert the ASR data into speech to text translation data using an MT system, and then generate the target audio from the target text with a TTS system. Unfortunately, these methods do not guarantee the accuracy of the generated pseudo S2ST data. Another line of work aims at boosting the performance of direct S2ST through pre-training methods [3, 9]. For example, the paper in [9] explores pre-training the encoder with mSLAM objective [10], and pre-training the decoder of Translatoron 2 [11] with MT task to generate phonemes. The authors in [3] propose to combine wav2vec 2.0 [12] encoder and mBART [13] decoder to a speech-to-unit translation (S2UT) model, which also can be further boosted by data augmentation techniques.

Although the self-supervised pre-training method in [3] can initialize the direct S2ST model with the pre-trained wav2vec 2.0 encoder and mBART decoder, which are trained with discrete hidden units [14] from unlabeled speech data. it still lacks an effective connection between encoder and decoder, and ignores the cross-lingual modeling capacity in pre-training. In the real world, speech data, ASR data, and MT data are relatively much more than direct S2ST corpora, and MT data can be utilized to learn the transformation ability from source text to target text. How to build the cross-lingual bridge between speech encoder and unit decoder of direct S2ST with bilingual text in the pre-training stage is not well explored.

In this paper, we propose a Speech2S model, which aims at modeling cross-lingual information and alleviating data scarcity problems by jointly pre-training with unpaired speech and bilingual MT text for the direct speech to speech translation task. More specially, Speech2S consists of a speech encoder, unit encoder, and unit decoder. We propose two pre-training tasks to pre-train the three modules with unit encoder as the bridge between source speech and target units. Like HuBERT [14], the first pre-training objective is to predict the clustered units based on the output of both speech encoder and unit encoder, with unlabeled speech data. To take advantage of bilingual machine translation corpus, we first leverage two text-to-unit models to convert source/target text into source/target units, with which, the cross-lingual unit encoder and decoder can be well pre-trained through cross-entropy loss.

We evaluate the proposed model on Europarl-ST [15] and VoxPopuli [5] S2ST datasets. Our contributions can be summarized as follows. (1) We propose a joint pre-trained Speech2S model, which can take advantage of bilingual text data to boost bilingual speech conversion. (2) The proposed model achieves a significant improvement of about 5 BLEU scores compared to the pre-trained model without MT data. (3) Furthermore, we conduct a detailed analysis about the effect of parallel data size, data augmentation of different domains, and subjective evaluation.

2. RELATED WORK

Conventional speech to speech translation is usually composed of cascaded ASR, MT and TTS modules [16, 17]. On this basis, to avoid error transmission caused by cascade models, researchers ex-

[†]Work done during internship at Microsoft Research Asia. Corresponding author: Long Zhou (lozhou@microsoft.com).

[1]Code and pre-trained models are available at `https://github.com/microsoft/SpeechT5/tree/main/Speech2S`.

plore the combination of ASR and MT modules [18, 19], as well as TTS modules [1, 20], namely direct S2ST. This paper focuses on exploring direct S2ST with improved pre-training methods.

2.1. Direct Speech to Speech Translation

S2ST, which directly translates the source speech to the target speech, has attracted a lot of attention recently [1, 2, 20–22]. Translatotron [1] is the first work to achieve direct speech-to-speech translation by using a sequence-to-sequence model. This system uses an encoder to model the log-mel spectrogram and predict the target spectrogram by the decoder, combined with the speaker information. Then, a vocoder is used to convert spectrogram into waveform. This work in [11] improves Translatotron system by utilzing a duration-based spectrogram synthesizer enhanced with target phoneme from decoder. Unlike Translatotron, the authors in [2] propose a novel direct speech to speech translation system, which employs discrete hidden units instead of spectrogram as model target before vocoder. They also expand it without using any text data on real-world S2ST tasks [23]. However, real speech to speech translation data is very limited due to the high cost of obtaining such data [5, 6]. Our work is to leverage a pre-training approach to alleviate data dependence on direct S2ST dataset.

2.2. Pre-Training for Direct S2ST

Recent years have witnessed a great progress on pre-training techniques for direct S2ST tasks [3, 9]. The work in [9] employs speech-text joint model from mSLAM as the encoder, to generate phoneme sequence with MT task and generate spectrogram with S2ST task. The most related work to our paper is [3], which enhances the speech-to-unit translation (S2UT) model by a wav2vec 2.0 [12] encoder and a decoder from pre-trained unit mBART [24]. In this S2UT model, wav2vec 2.0 is pre-trained on unlabeled audio data, and mBART leverages reduced discrete units tokenized from unlabeled audio data to train a denoised encoder-decoder model, and finally uses the mBART decoder to initialize the S2UT decoder. However, the simple combination of wav2vec 2.0 encoder and mBART decoder lacks cross-language modeling capabilities, which is particularly important for translation tasks. Motivated by this, we propose to bridge the language gap by utilizing machine translation corpus to improve model pre-training for direct speech to speech translation.

3. THE PROPOSED METHOD

Our goal is to leverage paired machine translation corpora to bridge the semantic gap between source speech and target speech. In this section, we will first introduce the model architecture of Speech2S, and the details of the model pre-training and fine-tuning methods.

3.1. Structure of Speech2S

As shown in Figure 1, Speech2S consists of a speech encoder $\mathcal{E}_s$, a unit encoder $\mathcal{E}_u$ and a unit decoder $\mathcal{D}_u$. Speech encoder and unit encoder employ standard Transformer network [25] with the same Transformer layers, except that a 5-layer CNN network in speech encoder is used to pre-process the original audio signal. Unit decoder is a multi-layer Transformer decoder layer which is composed of a multi-head self-attention mechanism, cross-attention mechanism, and a FFN network.

Formally, we denote unpaired speech as S, and denote bilingual text as (X, Y). After applying the speech and text discretization

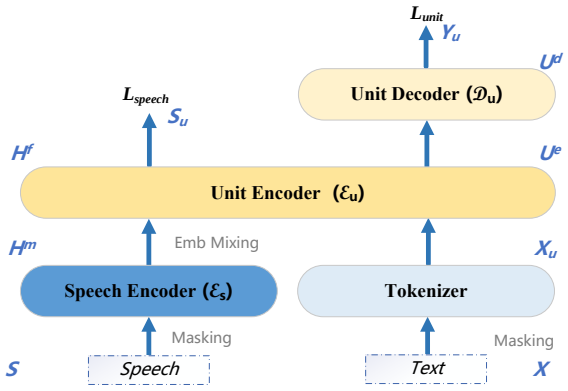

Fig. 1: The overall framework of the proposed Speech2S.

modules (as introduced in Section 3.2.1), we obtain the speech units S_u from S and bilingual units (X_u, Y_u) from (X, Y). Briefly speaking, $\mathcal{E}_s$ is used to encode the source audio sequence S into a sequence of vector representation H^m. Following the mixing mechanism proposed in [26], we also adapt it to improve alignment learning by randomly replacing part of H^m with the corresponding unit embedding. $\mathcal{E}_u$ can transform speech representation H^m into final hidden states H^f, or transform source unit sequence X_u into unit hidden states U^e. Besides, $\mathcal{D}_u$ reads the encoder representations and generates a target unit sequence Y_u.

3.2. Model Pre-Training

Before pre-training, we first use two discretization modules to tokenize speech and text into shared discrete tokens. Then the model can be optimized by two pre-training objectives, including speech to units task using speech encoder and unit encoder, and source units to target units task using unit encoder and unit decoder.

3.2.1. Speech/text discretization

We use HuBERT k-means cluster as the speech discretization module, which is learned from the HuBERT iter-1 hidden states, and can tokenize unlabeled speech into discrete hidden units. To tokenize text into the same space as speech, we introduce two text-to-unit models like [26] with the same model architecture, which are trained by using two small ASR corpus with paired speech and transcription. More specifically, we first use speech discretization to convert paired speech into hidden units, and obtain the ⟨text, unit⟩ data by combining it with paired text. Then we utilize a sequence-to-sequence model to achieve the text-to-unit models trained on the paired text and unit data. Once obtaining the discrete models, we can tokenize unlabeled speech S into hidden units S_u, and tokenize bilingual text (X, Y) into bilingual units (X_u, Y_u), respectively, all of which can be used to optimize the model in pre-training stage.

3.2.2. Pre-training objects

When the input audio S is fed into the speech encoder $\mathcal{E}_s$, it is partially masked and encoded into middle hidden states $H^m = \{h_1^m, h_2^m, ..., h_T^m\}$, namely $\mathcal{E}_s$(S), which also be sent to unit encoder $\mathcal{E}_u$ to get final hidden states $H^f = \{h_1^f, h_2^f, ..., h_T^f\}$ from $\mathcal{E}_u(H^m)$. Based on H^m and H^f, the speech pre-training object can be designed on the masked positions as,

$$\mathcal{L}_{speech} = -\sum_{t \in \mathcal{M}} \Big(\log p(u_t | h_t^m) + \log p(u_t | h_t^f) \Big) \quad (1)$$

where $u_t \in S_u$ is the target hidden units, and the $p(.)$ is parameterized as the same way with HuBERT [14].

Unit encoder $\mathcal{E}_u$ also takes X_u as input in pre-training stage, and use $\mathcal{E}_u(X_u)$ to output the encoded unit hidden state U^e. The unit decoder $\mathcal{D}_u$ will generate a series of hidden states $U^d = \mathcal{D}_u(U^e)$ according to the encoder representation of source units. The objective function of unit pre-training is formalized as,

$$\mathcal{L}_{unit} = -\sum_{i=0}^{Y_u} |\log p(y_{u,i}|Y_{u,<i}, U^e) \quad (2)$$

where $y_{u,i} \in Y_u$, $Y_{u,<i}$ denotes $\{y_{u,0}, y_{u,1}, ..., y_{u,i-1}\}$, and $p(.)$ is a softmax layer. Finally, we pre-train Speech2S under multi-task learning framework with $\mathcal{L} = \mathcal{L}_{speech} + \mathcal{L}_{unit}$.

3.3. Speech2S Fine-Tuning

In the fine-tuning stage, we can fine-tune Speech2S with speech encoder, unit encoder, and unit decoder to a direct speech-to-speech translation model. Leveraging the cross-entropy loss, we simply employ direct S2ST corpus as the fine-tuning dataset to optimize the model, where the target speech needs to convert into target units using speech discretization module. Finally, we utilize a unit-based HiFi-GAN [23] to generate the target waveform from target units.

4. EXPERIMENTS

4.1. Datasets

We conduct our experiments on two directions of the same language pair: Spanish-English (es-en) and English-Spanish (en-es). For pre-training, we use VoxPopuli dataset, a large-scale multilingual corpus providing 100K hours of unlabelled speech data in 23 languages, as speech pre-training data. The ASR subset of Voxpopuli (VoxPopuli-ASR) in each language is used to train the textual discretization module, namely sequence-to-sequence based text-to-unit model. We use machine translation data between English and Spanish from Europarl v10 [27] as the bilingual text data to generate paired text units for textual unit pre-training. Meanwhile, the speech-to-speech paired data VoxPopuli-S2S is used for our S2ST fine-tuning stage. We use the dev set split from VoxPopuli and the dev/test set of Europarl-ST dataset to verify the effect of speech to speech translation models. In order to avoid duplication with the corpus of the test set, we deleted the data of 2012 and earlier in the VoxPopuli training set. To avoid errors caused by audio itself, all audio is unified to the 16 kHz ogg format. In addition, we use the training sets text of CoVoST-2 and Europarl-ST datasets for additional analysis experiments on data augmentation for different domains. Data details are shown in Table 1.

Table 1: Statistics of datasets (train/dev/test splits), including pre-training, fine-tuning, and tokenizing datasets.

data	samples	source(hrs)	target(hrs)
pre-train, en-es			
VoxPopuli	1.8M	14k	-
Europarl v10	1.9M	-	-
pre-train, es-en			
VoxPopuli	2.0M	16k	-
Europarl v10	1.9M	-	-
fine-tune, en-es			
VoxPopuli-S2S	120k/6k/-	394/20/-	403/21/-
fine-tune, es-en			
VoxPopuli-S2S	153k/6k/-	513/19/-	495/18/-
Europarl-ST	31.6k/1.3k/1.3k	75.6/3.0/2.9	76.5/3.0/-
CoVoST-2	78.9k/13.3k/13.2k	112.0/22.0/22.7	81.0/14.4/-
tokenize, en			
VoxPopuli-ASR	-	1.3k	-
tokenize, es			
VoxPopuli-ASR	-	261	-

4.2. Implementation Details

Discretization We use released k-means cluster model[2] from multilingual HuBERT (mHuBERT), which trained with VoxPopuli 100k subset [23], to extract units from speech data. For text discretization, we first extract the units of Voxpopuli-ASR speech using mHuBERT cluster and normalize the units using the same 1h English or Spanish speech normalizer as [23]. Then we train the text-to-unit discretization model using the normalized units and transcripts of the corresponding speech of the units. The text-to-unit model has 6 Transformer layers as encoder and 6 layers for decoder, each has 512 nodes with 4 attention heads. Pairs of translation text in Europarl v10 are pre-extracted offline using this discretization model and the extracted units are applied in the pre-training stage.

Pre-training Our Speech2S is composed of a 6-layer Transformer speech encoder, a 6-layer Transformer unit encoder a 6-layer Transformer decoder and an output FFN layer of 1024 units. Each Transformer layer has 768 nodes with 4 attention heads and relative positional attention bias [28]. We pre-train with the same 400k training steps for all models.

Fin-tuning The fine-tuning model structure is basically the same as the pre-training model structure. The normalized units of target language used in fine-tuning stage are extracted using the same extractor as the text-to-unit model. After generating units, we use unit-based HiFi-GAN [23] to generate target speech. English and Spanish use recognition models wav2vec[3] and microsoft speech-to-text tookit[4] to transcribe into text, respectively. The SacreBLEU toolkit [29] is used to calculate the final BLEU score.

Baselines For comparison, we design two strong baselines for the experiment. The first one employs HuBERT encoder to initialize the encoder of speech-to-unit translation model, and the other is existing S2UT model [3], which is initialized with HuBERT encoder plus 6-layer unit level mBART decoder. The two models use the same speech data as our model for pre-training and fine-tuning. The parameters of the S2UT base model and our Speech2S model are almost the same.

4.3. Experimental Results

Table 2 shows the BLEU scores of S2UT systems [3] and our Speech2S systems. By comparing the model fine-tuned from HuBERT and our proposed model, results show that our model achieves more than 4 BLEU value gains on the S2ST tasks in both directions (#5 vs. #3). Compared to S2UT base model fine-tuned from HuBERT encoder and mBART decoder, the proposed Speech2S model still has an improvement of more than 3 BLEU scores (#5 vs. #4). This result proves that our model can better incorporate text information into the language model through pre-training, and learn the corresponding relationship between source language speech and

[2]https://github.com/facebookresearch/fairseq/blob/main/examples/speech_to_speech/docs/textless_s2st_real_data.md

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

[4]https://azure.microsoft.com/zh-cn/products/cognitive-services/speech-to-text

Table 2: Speech to speech translation performance (BLEU) on VoxPopuli dev set and Europarl-ST dev/test sets. For the S2UT systems, the results on VoxPopuli are reproduced by ourselves, and the results of Europarl-ST are reported in the paper.

#	System	Pre-trained Model	Parameters	en-es		es-en	
				VoxPopuli	Europarl-ST	VoxPopuli	Europarl-ST
1	S2UT [3]	w/o pre-training	Large (827M)	-	-/21.8	-	-/18.8
2		wav2vec 2.0+mBART		24.3	**25.7/26.0**	21.4	25.7/23.8
3	Ours	HuBERT	Base (157M)	20.5	20.2/19.1	18.7	21.1/19.2
4		HuBERT+mBART [3]		22.5	21.8/20.9	20.1	23.2/21.1
5		Speech2S		**24.6**	25.3/25.6	**23.3**	**26.8/24.4**

target language units through shared unit encoder. Furthermore, we compare our model with S2UT Large model from their paper (#5 vs. #2), our method achieves almost the same results as S2UT Large on the English-Spanish task with a smaller number of parameters, while on the Spanish-English test set, it achieves results that exceed those of the larger model, which also verifies the above conclusion.

4.4. Analysis

4.4.1. Effect of Parallel Data Size

An interesting question is how well does the model perform if we only have very little fine-tuning data. Here, we verify the effect of varying parallel data size for Speech2S and baselines. We evaluate the proposed Speech2S and baseline from HuBERT on 10 hour, 50 hour, and 100 hour supervised data sets respectively. These training data are randomly sampled from all data of VoxPopuli-S2S.

Table 3: BLEU scores for Speech2S and baseline trained with 15-hr, 50-hr, and 100-hr subsets.

Pre-trained Model	hours	en-es		es-en	
		dev	test	dev	test
HuBERT	10	0.3	0.5	0.5	0.5
Speech2S (Ours)	10	12.3	11.9	20.1	19.4
HuBERT	50	10.2	11.2	12.6	12.9
Speech2S (Ours)	50	19.4	18.8	26.8	24.4
HuBERT	100	12.9	13.7	15.7	14.1
Speech2S (Ours)	100	23.2	23.5	24.6	23.1

From Table 3, we can find that even if there is only 10 hours of supervised data, through our joint pre-training with speech and bilingual text, the BLEU can reach more than 10. On the 100 hour supervised data set, the fine-tuning results are close to those of hundreds of hours of supervised data fine-tuning. From the results of weak supervision, we can draw a conclusion that the Speech2S model can learn the unified mapping of speech and unit well through pre-training, thus reducing the dependence on supervised S2ST data.

4.4.2. Effect of Data Augmentation

In this section, we explore the effect of data augmentation for different domain datasets. As shown in Table 4, we first evaluate the performance on CoVoST-2 dev/test sets using the model trained with VoxPoluli train set. In terms of absolute performance, the BLEU scores of CoVoST-2 underperform significantly that of Europarl-ST (#3 vs. #1). A potential reason is that the pre-training and fine-tuning data domains are consistent for Europarl-ST test set, but it has a domain mismatch problem between VoxPopuli and CoVoST-2.

We conduct data augmentation experiments by adding the paired source speech and target unit data from Europral-ST and CoVoST-2 speech-to-text translation dataset. Based on the training data, which consists of source speech and target text, we use the text-to-unit model trained on VoxPopuli-ASR data to convert the text of the target language into units, and then enlarge the training set with the speech and generated target units, as shown in the line 2 and 4 of Table 4. With data augmentation, the Speech2S can achieve bigger improvements on CoVoST-2 than Europarl-ST, which confirms our suspicions. Experimental results also demonstrate that this data augmentation method is very effective for domain adaption.

Table 4: BLEU scores with data augmentation for different domain datasets. *vp_train* means the VoxPopuli training set, *Eur_train* means the Europarl-ST training set, and *Cov_train* means the CoVoST-2 training set.

#	Fine-tuning Data	Evaluation Data	dev	test
1	*vp_train*	Europarl-ST	26.8	24.4
2	*vp_train+Eur_train*		**29.3**	**26.1**
3	*vp_train*	CoVoST-2	15.7	17.6
4	*vp_train+Cov_train*		**24.2**	**26.9**

4.4.3. Subjective Evaluation

To further compare the speech quality generated by different models, we select 50 samples from the Europarl-ST dev set and test the naturalness score of these samples. Table 5 lists the naturalness score of different models, including S2UT model and our Speech2S models without and with data augmentation. The results show that our proposed Speech2S achieves the naturalness score of 4.1, outperforming S2UT model fine-tuned from HuBERT and mBART. With data augmentation, the Speech2S model obtains the best naturalness score of 4.3. Experiments demonstrate that our proposed method not only significantly improves the translation quality of S2ST tasks, but also enhances the naturalness of generated speech. In addition, we can find from this experiment that more accurate units will also help to improve the quality of the final synthesized speech.

Table 5: The naturalness score for different models. DAT means data augmentation method.

Model	S2UT	Speech2S	Speech2S+DAT
naturalness score	4.0±0.1	4.1±0.1	4.3±0.1

5. CONCLUSION

This paper proposes a novel pre-training method with unlabeled speech and paired text data for direct speech to speech translation. The core of the proposed Speech2S is to enhance the cross-lingual speech conversion capability by modeling the transformation from source units to target units, which are extracted from bilingual text data using a discrete tokenizer. Experimental results and analyses on common VoxPopuli and Europarl-ST speech-to-speech translation tasks demonstrate the effectiveness and superiority of the proposed Speech2S model.

6. REFERENCES

[1] Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," *arXiv preprint arXiv:1904.06037*, 2019.

[2] Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, et al., "Direct speech-to-speech translation with discrete units," *arXiv preprint arXiv:2107.05604*, 2021.

[3] Sravya Popuri, Peng-Jen Chen, Changhan Wang, Juan Pino, Yossi Adi, Jiatao Gu, Wei-Ning Hsu, and Ann Lee, "Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation," *arXiv preprint arXiv:2204.02967*, 2022.

[4] Changhan Wang, Hirofumi Inaguma, Peng-Jen Chen, Ilia Kulikov, Yun Tang, Wei-Ning Hsu, Michael Auli, and Juan Pino, "Simple and effective unsupervised speech translation," *arXiv preprint arXiv:2210.10191*, 2022.

[5] Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, et al., "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *ACL*, Aug. 2021, pp. 993–1003.

[6] Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen, "Cvss corpus and massively multilingual speech-to-speech translation," *arXiv preprint arXiv:2201.03713*, 2022.

[7] Paul-Ambroise Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswani, et al., "SpeechMatrix: A large-scale mined corpus of multilingual speech-to-speech translations," .

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

[9] Ye Jia, Yifan Ding, Ankur Bapna, Colin Cherry, Yu Zhang, Alexis Conneau, and Nobuyuki Morioka, "Leveraging unsupervised and weakly-supervised data to improve direct speech-to-speech translation," *arXiv preprint arXiv:2203.13339*, 2022.

[10] Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau, "mslam: Massively multilingual joint pre-training for speech and text," *arXiv preprint arXiv:2202.01374*, 2022.

[11] Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz, "Translatotron 2: Robust direct speech-to-speech translation," *arXiv preprint arXiv:2107.08661*, 2021.

[12] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[13] Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli, "Multilingual speech translation with efficient finetuning of pretrained models," *arXiv preprint arXiv:2010.12829*, 2020.

[14] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman 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.

[15] Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerda, Javier Jorge, Nahuel Roselló, Adria Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan, "Europarl-st: A multilingual corpus for speech translation of parliamentary debates," in *ICASSP*. IEEE, 2020, pp. 8229–8233.

[16] Satoshi Nakamura, Konstantin Markov, Hiromi Nakaiwa, Genichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, J-S Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto, "The atr multilingual speech-to-speech translation system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 365–376, 2006.

[17] Alon Lavie, Alex Waibel, Lori Levin, Michael Finke, Donna Gates, Marsal Gavalda, Torsten Zeppenfeld, and Puming Zhan, "Janus-iii: Speech-to-speech translation in multiple languages," in *ICASSP*. IEEE, 1997, vol. 1, pp. 99–102.

[18] Evgeny Matusov, Stephan Kanthak, and Hermann Ney, "On the integration of speech recognition and statistical machine translation," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[19] Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," *arXiv preprint arXiv:1612.01744*, 2016.

[20] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura, "Speech-to-speech translation between untranscribed unknown languages," in *ASRU*. IEEE, 2019, pp. 593–600.

[21] Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura, "Transformer-based direct speech-to-speech translation with transcoder," in *SLT*. IEEE, 2021, pp. 958–965.

[22] Chen Zhang, Xu Tan, Yi Ren, Tao Qin, Kejun Zhang, and Tie-Yan Liu, "Uwspeech: Speech to speech translation for unwritten languages," in *AAAI*, 2021, vol. 35, pp. 14319–14327.

[23] Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Juan Pino, Jiatao Gu, and Wei-Ning Hsu, "Textless speech-to-speech translation on real data," *arXiv preprint arXiv:2112.08352*, 2021.

[24] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[26] Ziqiang Zhang, Long Zhou, Junyi Ao, Shujie Liu, Lirong Dai, Jinyu Li, and Furu Wei, "Speechut: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training," *arXiv preprint arXiv:2210.03730*, 2022.

[27] Philipp Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of machine translation summit x: papers*, 2005, pp. 79–86.

[28] Xing Wang, Zhaopeng Tu, Longyue Wang, and Shuming Shi, "Self-attention with structural position representations," *arXiv preprint arXiv:1909.00383*, 2019.

[29] Matt Post, "A call for clarity in reporting bleu scores," *arXiv preprint arXiv:1804.08771*, 2018.