INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

SpecRec: An Alternative Solution for Improving End-to-End Speech-to-Text Translation via Spectrogram Reconstruction

Junkun Chen[‡§*] , *Mingbo Ma*[†*], *Renjie Zheng*[†], *Liang Huang*[†‡]

[‡]Oregon State University, Corvallis, OR, USA
[†]Baidu Research, Sunnyvale, CA, USA
chenjun2@oregonstate.edu, mingboma@baidu.com

Abstract

End-to-end Speech-to-text Translation (E2E-ST), which directly translates source language speech to target language text, is widely useful in practice, but traditional cascaded approaches (ASR+MT) often suffer from error propagation in the pipeline. On the other hand, existing end-to-end solutions heavily depend on the source language transcriptions for pre-training or multi-task training with Automatic Speech Recognition (ASR). We instead propose a simple technique to learn a robust speech encoder in a self-supervised fashion only on the speech side, which can utilize speech data without transcription. This technique termed Spectrogram Reconstruction (SpecRec), learns better speech representation via recovering the missing speech frames and provides an alternative solution to improving E2E-ST. We conduct our experiments over 8 different translation directions. In the setting without using any transcriptions, our technique achieves an average improvement of +1.1 BLEU. SpecRec also improves the translation accuracy with +0.7 BLEU over the baseline in speech translation with ASR multitask training setting.

Index Terms: speech translation, self-reconstruction

1. Introduction

Speech-to-text translation (ST), which translates the source language speech to target language text, is useful in many scenarios such as international conferences, travels, foreign-language video subtitling, etc. Conventional cascaded approaches to ST [1, 2, 3, 4] first transcribe the speech audio into source language text (ASR) and then perform text-to-text machine translation (MT), which inevitably suffers from error propagation in the pipeline. To alleviate this problem, recent efforts explore end-to-end approaches (E2E-ST) [5, 6, 7, 8], which are computationally more efficient at inference time and mitigate the risk of error propagation from imperfect ASR.

To improve the translation accuracy of E2E-ST models, researchers either initialize the encoder of ST with a pre-trained ASR encoder [6, 9, 10] to get better representations of the speech signal, or perform Multi-Task Learning (MTL) with ASR to bring more training and supervision signals to the shared encoder [11, 12, 13, 14] (see Fig. 1). These methods improve the translation quality by providing more training signals to the encoder to learn better phonetic information and hidden representation correspondence [15].

However, both above solutions assume the existence of substantial speech transcriptions of the source language. Unfortunately, this assumption is problematic. On the one hand, for certain low-resource languages, especially endangered ones [16, 17], the source speech transcriptions are expensive to collect. Moreover, according to the report from [18], there are more than 3000 languages that have no written form or no standard orthography, making phonetic transcription impossible [19]. On the other hand, the amount of speech audios with transcriptions are limited (as they are expensive to collect), and there exist far more audios without any annotations. It will be much more straightforward and cheaper to leverage these raw audios to train a robust encoder directly.

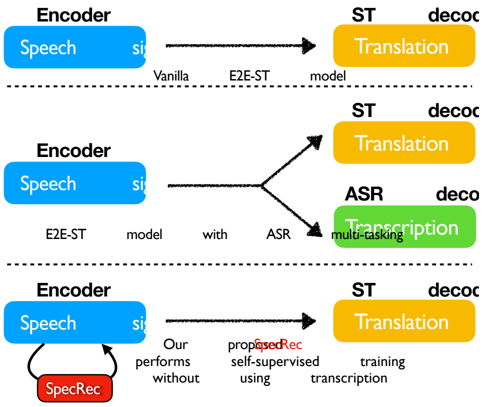

Figure 1: *Comparisons with different existing solutions and our proposed Spectrogram Reconstruction (SpecRec).*

To relieve from the dependency on source language transcriptions, we present a straightforward yet effective solution, Spectrogram Reconstruction (SpecRec), to utilize the speech data in a self-supervised fashion without using any source language transcription, unlike other speech translation models that rely on source speech transcription via ASR pre-training or ASR MTL. Aside from the regular training of E2E-ST (without ASR as MTL or pre-training), SpecRec masks certain portions of the speech input randomly and aims to recover the masked speech signals with their context on the encoder side. The contributions of our paper are as follows:

- We demonstrate the importance of a self-supervising module for E2E-ST. Unlike all previous attempts, which heavily depend on transcription, SpecRec improves the capacity of the encoder by recovering masked speech signals merely based on their context. SpecRec also can be used together with transcriptions in ASR pre-training and MTL settings to further boost the translation accuracy.
- For 8 different translation directions, when we do not use any transcription, SpecRec demonstrates an average BLEU improvement of 1.1 in the basic setting and 0.7 in ST with the ASR MTL setting.

§ work conducted during internship at Baidu Research
* equal contribution

http://dx.doi.org/10.21437/Interspeech.2021-733

2. Preliminaries: ASR and ST

We first briefly review the standard E2E-ST and E2E-ST with ASR MTL to set up the notations.

2.1. Vanilla direct E2E-ST Training with Seq2Seq

Regardless of particular design of Seq2Seq models for different tasks, the encoder always takes the source input sequence $\boldsymbol{x} = (x_1, ..., x_n)$ of n elements where each $x_i \in \mathbb{R}^{d_x}$ is a d_x-dimension vector and produces a sequence of hidden representations $\boldsymbol{h} = f(\boldsymbol{x}) = (h_1, ..., h_n)$ where $h_i = f(\boldsymbol{x})$. The encoding function f can be implemented by a combination between Convolution, RNN and Transformer. More specifically, $\boldsymbol{x}$ can be the spectrogram or mel-spectrogram of the source speech, and each x_i represents the frame-level speech feature with certain duration.

On the other hand, the decoder greedily predicts a new output word y_t given both the source sequence $\boldsymbol{x}$ and the prefix of decoded tokens, denoted $\boldsymbol{y}_{<t} = (y_1, ..., y_{t-1})$. The decoder continues the generation until it emits `<eos>` and finishes the entire decoding process. Finally, we obtain the hypothesis $\boldsymbol{y} = (y_1, ..., \texttt{<eos>})$ with the model score which defined as following:

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{t=1}^{|\boldsymbol{y}|} p(y_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}) \tag{1}$$

During the training time, the entire model aims to maximize the conditional probability of each ground-truth target sentence $\boldsymbol{y}^\star$ given input $\boldsymbol{x}$ over the entire training corpus $D_{\boldsymbol{x},\boldsymbol{y}^\star}$, or equivalently minimizing the following loss:

$$\ell_{\text{ST}}(D_{\boldsymbol{x},\boldsymbol{y}^\star}) = -\sum_{(\boldsymbol{x},\boldsymbol{y}^\star)\in D_{\boldsymbol{x},\boldsymbol{y}^\star}} \log p(\boldsymbol{y}^\star \mid \boldsymbol{x}) \tag{2}$$

2.2. Multi-task Learning with ASR

To further boost the performance of E2E-ST, researchers proposed to either use a pre-trained ASR encoder to initialize ST encoder, or to perform ASR MTL together with ST training. We only discuss the MTL since pre-training does not require significant change to the Seq2Seq model.

During multi-task training, there are two decoders sharing one encoder. Besides the MT decoder, there is also another decoder for generating transcriptions. With the help of ASR training, the encoder is able to learn more accurate speech segmentations (similar to forced alignment) making the global reordering of those segments for MT relatively easier. We defined the following training loss for ASR:

$$\ell_{\text{ASR}}(D_{\boldsymbol{x},\mathbf{z}^\star}) = -\sum_{(\boldsymbol{x},\mathbf{z}^\star)\in D_{\boldsymbol{x},\mathbf{z}^\star}} \log p(\mathbf{z}^\star \mid \boldsymbol{x}) \tag{3}$$

where $\mathbf{z}^\star$ represents the annotated, ground-truth transcription for speech audio $\boldsymbol{x}$. In our baseline setting, we also use a hybrid CTC/Attention framework [20] on the encoder side. In the case of joint training with ASR for ST, the total loss is defined as

$$\ell_{\text{MTL}}(D_{\boldsymbol{x},\boldsymbol{y}^\star,\mathbf{z}^\star}) = \ell_{\text{ST}}(D_{\boldsymbol{x},\boldsymbol{y}^\star}) + \ell_{\text{ASR}}(D_{\boldsymbol{x},\mathbf{z}^\star}) \tag{4}$$

where $D_{\boldsymbol{x},\boldsymbol{y}^\star,\mathbf{z}^\star}$ is the training dataset which contains speech, translation and transcription triplets.

3. E2E-ST Training with SpecRec

All the existing solutions to boost the current E2E-ST performance heavily depend on the availability of the transcription of the source language. Those solutions are not able to take advantage of the large amount of speeches without any annotations. They also become inapplicable when the source language is low-resource or even does not have a standard orthography system. Therefore, the ideal solution should not be constrained by source language transcription and still achieves similar translation quality. Thus, we introduce SpecRec in this section.

3.1. SpecRec as Part of Training Objective

We propose to perform self-supervised training on the encoder side by reconstructing sabotaged speech signals from the input. Note that SpecRec is totally different from another self-supervised training [21, 22, 23] which relies on transcription to segment the speech audio with forced alignment tools[24, 25]. We directly apply random masks with different widths over speech audio, eliminating the dependency of transcription. Therefore, SpecRec can be easily applied to speech audio without transcription and even to any non-human speech audio, e.g., music and animal sound.

Formally, we define a random replacement function over the original speech input $\boldsymbol{x}$:

$$\hat{\boldsymbol{x}} \sim \text{Mask}_{\text{frame}}(\boldsymbol{x}, \lambda), \tag{5}$$

where $\text{Mask}(\cdot)_{\text{frame}}$ randomly replaces some certain frames in $\boldsymbol{x}$ with the same random initialized vector, $\epsilon \in \mathbb{R}^{d_x}$, with a probability of λ (30% in our experiments). Note that we use the same vector ϵ to represent all the corrupted frames. Then we obtain a corrupted input $\hat{\boldsymbol{x}}$ and its corresponding latent representation $\hat{\boldsymbol{h}}$.

For the the SpecRec module, we have the following training objective to reconstruct the original speech signal with the surrounding context information with self-supervised fashion:

$$\ell_{\text{Rec}}(D_{\boldsymbol{x}}) = \sum_{\boldsymbol{x}\in D_{\boldsymbol{x}}} ||\boldsymbol{x} - \phi(f(\hat{\boldsymbol{x}}))||_2^2 \tag{6}$$

where ϕ is a reconstruction function which tries to recover the original signal from the hidden representation $f(\hat{\boldsymbol{x}})$ with corrupted inputs. For simplicity, we use regular 2D deconvolution as ϕ, and mean squared error for measuring the difference between original input and recovered signal. Finally, we have the following total loss of our model

$$\ell_{\text{SpecRec}}(D_{\boldsymbol{x},\boldsymbol{y}^\star}) = \ell_{\text{ST}}(D_{\boldsymbol{x},\boldsymbol{y}^\star}) + \ell_{\text{Rec}}(D_{\boldsymbol{x}})$$

To further boost the performance of E2E-ST, we can train SpecRec with ASR MTL when transcription is available:

$$\ell_{\text{SpecRec + MTL}}(D_{\boldsymbol{x},\boldsymbol{y}^\star,\mathbf{z}^\star}) = \ell_{\text{MTL}}(D_{\boldsymbol{x},\boldsymbol{y}^\star,\mathbf{z}^\star}) + \ell_{\text{Rec}}(D_{\boldsymbol{x}})$$

3.2. Different Masking Strategies

SpecRec aims at much harder tasks than pure textual pre-training models, e.g., BERT or ERINE, which only perform semantic learning over missing tokens. In our case, we not only try to recover semantic meaning, but also acoustic characteristics of given audio. SpecRec simultaneously predicts the missing words and generates spectrograms like speech synthesis tasks.

To ensure the masked segments contain different levels of granularity of speech semantics, we propose the following masking methods. **Single Frame Masking** Uniformly mask λ% frames out of $\boldsymbol{x}$ to construct $\hat{\boldsymbol{x}}$. Note that we might have continuous frames that were masked. **Span Masking** Similar with SpanBERT [26], we first sample a serial of span widths and then apply those spans randomly to different positions of the input signal. Note that we do not allow overlap in this case. Our span masking is defined as $\hat{\boldsymbol{x}} \sim \text{Mask}_{\text{span}}(\boldsymbol{x}, \lambda)$.

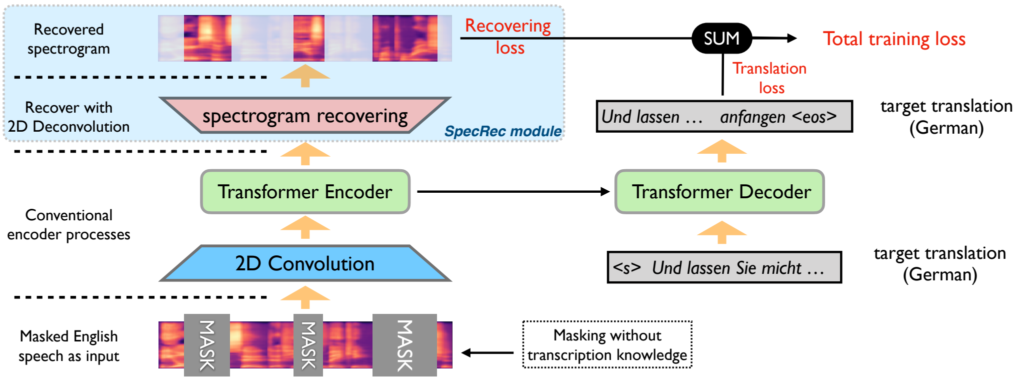
Recovered spectrogram
Recovering loss
SUM
Total training loss
Translation loss
Recover with 2D Deconvolution
spectrogram recovering
SpecRec module
Und lassen … anfangen <eos>
target translation (German)
Transformer Encoder
Transformer Decoder
Conventional encoder processes
2D Convolution
<s> Und lassen Sie micht …
target translation (German)
Masked English speech as input
MASK
MASK
MASK
Masking without transcription knowledge

Figure 2: *SpecRec (in blue box) can be treated as one extra module besides standard Transformer encoder-decoder and convolution layers for processing speech signals.*

Table 1: *SpecRec only has 6.5% more parameters than the baseline model while ASR multi-tasking needs to use 51.6% more parameters.*

	ST	ST+ASR	ST+SpecRec
# of parameters	31M	47M	33M

4. Experiments

In this section, we analyze the performance of SpecRec in E2E-ST with 8 different language translation directions using English as the source speech on MuST-C dataset [27]. All raw audio files are processed by Kaldi [24] to extract 80-dimensional log-Mel filterbanks stacked with 3-dimensional pitch features using a window size of 25 ms and step size of 10 ms. We train sentencepiece [28] models with a joint vocabulary size of 8K for each dataset. We remove samples that have more than 3000 frames for GPU efficiency. Our basic Transformer based E2E-ST framework has similar settings with ESPnet-ST[29]. We first downsample the speech input with 2 layers of 2D convolution of size 3 with stride size of 2. Then there is a standard 12-layers Transformer with 2048 hidden size to bridge the source and target side. We only use 4 attention heads on each side of the transformer and each of them has a dimensionality of 256. For the SpecRec module, we simply linearly project the outputs of the Transformer encoder to another latent space, then upsample the latent representation with 2-layers deconvolution to match the size of the original input signal. For the random masking ratio λ, we choose 30% across all the experiments. During inference, we do not perform any masking over the speech input. We average the last 5 checkpoints for testing. For decoding, we use a beam search with setting beam size and length penalty to 5 and 0.6, respectively. We report detokenized case-sensitive BLEU.

Our SpecRec is very easy to replicate as we do not perform any parameters and architecture search upon the baseline system. Due to the simple, but effective design of SpecRec, SpecRec does not rely on intensive computation. It converges within 2 days of training with 8 1080Ti GPUs for the basic model. It is much faster than ASR MTL, or two stage training that the model initialized from a pretrained ASR model. We showcase the comparison of parameters between different solutions to E2E-ST in Table. 1

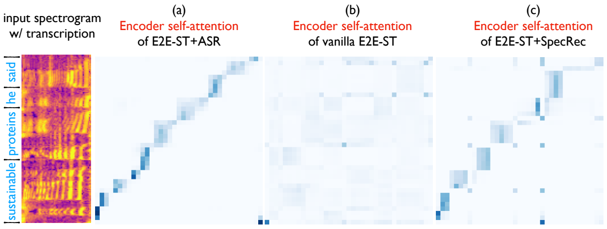

Figure 3: *One head of the last layer self-attention comparison between different models. ASR MTL and SpecRec help the encoder learns similar self-attentions. See discussion in 4.1.*

4.1. Analyzing ASR and SpecRec

Aside from the extra training signal that is introduced by transcriptions, there is a deeper reason why ASR and SpecRec are beneficial to E2E-ST. In this section, we first discuss the difficulties and challenges in E2E-ST. Then we analyze the reasons why ASR MTL and SpecRec are helpful for E2E-ST by visualizing the self-attention over source side encoder.

Compared with other tasks, e.g., MT or ASR, which also employ Seq2Seq framework for E2E training, E2E-ST is a more challenging task in many ways. Firstly, data modalities are different on the source and target sides. For ST, the encoder deals with speech signals and tries to learn word presentations on the decoder side, while MT has text format on both sides. Secondly, due to the nature of the high sampling rate of speech signals, speech inputs are generally multiple (e.g. 4 to 7) times longer than the target sequence, which increases the difficulties of learning the correspondence between source and target. Thirdly, compared with the natural monotonicity of the alignment of ASR, ST usually needs to learn the global reordering between speech signal and translation, and this raises the difficulties to another level. Especially in ST, since source and target are in different languages, it is very challenging to obtain the corresponding phoneme or syllable segments given the training signal from a different language.

Fig. 3 tries to explain and analyze the difference between E2E-ST (a) and E2E-ST with ASR MTL (b). We extract the topmost layer from the encoder for comparison. We notice that E2E-ST (a) tends to get more meaningful self-attention on the encoder with the training signal from ASR. With help from ASR, the source input spectrogram is chunked into segments

Table 2: *Comparisons between SpecRec and other baselines over 8 languages on MuST-C. We use SpecRec as an extra training module for E2E-ST. We notice that SpecRec with span masking achieves better performance and there is 1.09 BLEU score improvements upon E2E-ST. The column starting with "Avg. Δ" summarizes the average improvements upon the baseline method, E2E-ST.*

	Models	De	Es	Fr	It	Nl	Pt	Ro	Ru	Avg. Δ
Baselines	MT with ASR annotation [27]	28.09	34.16	42.23	30.40	33.43	32.44	28.16	18.30	
	Cascaded methods [29]	23.65	28.68	33.84	24.04	27.91	29.04	22.68	16.39	
	E2E-ST	19.64	23.68	28.91	19.95	23.01	24.00	21.06	12.05	-
	E2E-ST + SpecAug	20.06	24.51	29.26	20.27	23.73	24.40	21.21	12.84	+0.49
Ours	E2E-ST + SpecRec (single)	20.34	24.46	29.18	19.52	23.81	24.56	21.37	12.57	+0.44
	E2E-ST + SpecRec (span)	20.78	25.34	30.26	20.51	24.46	24.90	21.62	13.14	**+1.09**

Table 3: *Comparisons between SpecRec with ASR MTL and E2E-ST with ASR MTL. SpecRec still achieves an improvement about +0.7 BLEU. The column starting with "Avg. Δ" summarizes the average improvements upon the baseline method.*

Models	De	Es	Fr	It	Nl	Pt	Ro	Ru	Avg. Δ
E2E-ST+ASR MTL	21.70	26.83	31.36	21.45	25.44	26.52	23.71	14.54	-
E2E-ST+ASR MTL+SpecRec	22.41	26.89	32.55	22.12	26.49	27.22	24.45	14.90	+0.69

that contain phoneme-level information. During training, the natural monotonicity of the ASR alignment functions as a forced alignment to group a set of adjacent frames to represent certain phonemes or syllables from source speech. With a larger scale of segmented spectrograms, the target side decoder only needs to perform reordering on those segments instead of frames. Our observations also align with the analysis from [15].

We also visualize the self-attention on encoder for E2E-ST with SpecRec in (c) of Fig. 3. We find that SpecRec has the similar ability with ASR to segment the source speech into chunks. Recovering local frames usually needs surrounding contextual information, especially for the speaker and environment-related characteristic. But we still observe that self-attention sometimes focuses on longer distance frames as well. This type of attention is very similar with low to mid layer self-attention of ASR.

To conclude, we observe that SpecRec functions very similar to ASR on the encoder side. Hence, SpecRec is a reliable framework that can be used as an alternative solution when there is no transcription available.

4.2. Translation Accuracy Comparisons

We showcase the translation accuracy of SpecRec comparing against 6 baselines from Table 2 to Table 3:

- **MT with ASR annotation**: an MT system which generates the target translation from the gold transcription.
- **Cascaded**: cascade framework first transcribes the speech into transcription then passes the results to later machines translation system.
- **E2E-ST**: this is the basic direct translation system which does not use transcriptions in MuST-C.
- **E2E-ST + SpecAugment**: a data augmentation method [30, 31] by performing augmentation on the spectrogram.
- **E2E-ST + ASR MTL**: ST trained with ASR MTL using the transcription in MuST-C.

To better make a conclusion of our results from Table 2 to Table 3, we organize the comparisons as follows.

In Table 2, we first compare SpecRec against E2E-ST where there is no transcription for training. Both SpecRec with single and span masking methods achieve averagely +0.44 (single) and +1.09 (span) improvements in BLEU score correspondingly against E2E-ST in 8 different translation directions. Span masking consistently outperforms single frame masking as it is a more difficult self-supervised task.

In Table 3, we do the comparison where transcription is available for training. In this setting, we use "E2E-ST + ASR MTL" as the baseline. Joint training with SpecRec achieves +0.7 average improvements over 8 languages.

5. Related Work

Text-based BERT-style [32, 33, 34, 35] frameworks are extremely popular in recent years due to the remarkable improvement on the downstream tasks at fine-tuning stages. Inspired by techniques mentioned above, SpecRec also performs self-supervised training that masks certain portions randomly over the input signals. But different from BERT-style pre-training, SpecRec tries to recover the missing semantic information (e.g., words, subword units) and learns the capabilities to restore the missing speech characteristics and generate the original speech.

SpecAugment [30] was originally proposed as a data augmentation method by applying dropout over input speech, then it is adapted to ST by [31]. However, there is no recovering step in SpecAugment, and it cannot be jointly optimized.

[36, 37] use an additional module for pretrained ASR decoder to bridge the modality between speech and translation. [21, 23, 22] proposed to use forced-alignment to segment speech audio into pieces at word level and masked some certain words during fine-tuning. Obviously those approaches rely on the transcriptions of source speech, and cannot be applied to corpora without transcription. FAT-MLM [38] extends text-based MLM with a spectrogram reconstruction method to do speech and text multi-modal pretraining.

6. Conclusions

We have presented a novel alternative solution, SpecRec in this paper to improve the E2E-ST. We demonstrate the effectiveness of SpecRec with multiple different experiment settings in 8 languages. SpecRec is a simple but effective component for E2E-ST especially when there is no speech transcription available.

7. References

[1] H. Ney, "Speech translation: coupling of recognition and translation," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, vol. 1, pp. 517–520 vol.1.

[2] E. Matusov, S. Kanthak, and H. Ney, "On the integration of speech recognition and statistical machine translation," in *INTERSPEECH*, 2005.

[3] L. Mathias and W. Byrne, "Statistical phrase-based speech translation," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, 2006.

[4] A. Berard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," *CoRR*, vol. abs/1612.01744, 2016. [Online]. Available: http://arxiv.org/abs/1612.01744

[5] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," in *Proc. Interspeech*, 2017.

[6] A. Berard, L. Besacier, A. Kocabiyikoglu, and O. Pietquin, "End-to-end automatic speech translation of audiobooks," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6224–6228, 2018.

[7] L. C. Vila, C. Escolano, J. A. R. Fonollosa, and M. R. Costa-jussà, "End-to-end speech translation with the transformer," in *IberSPEECH*, 2018.

[8] M. A. D. Gangi, M. Negri, and M. Turchi, "Adapting Transformer to End-to-End Spoken Language Translation," in *Proc. Interspeech*, 2019.

[9] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," in *Proceedings of the 2019 Conference of NAACL-HLT Volume 1 (Long and Short Papers)*, 2019.

[10] C. Wang, Y. Wu, S. Liu, Z. Yang, and M. Zhou, "Bridging the gap between pre-training and fine-tuning for end-to-end speech translation," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 2020.

[11] A. Anastasopoulos, D. Chiang, and L. Duong, "An unsupervised probability model for speech-to-translation alignment of low-resource languages," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

[12] A. Anastasopoulos and D. Chiang, "Tied multitask learning for neural speech translation," in *Proceedings of the 2018 Conference of NAACL-HLT, Volume 1 (Long Papers)*, 2018.

[13] M. Sperber, G. Neubig, J. Niehues, and A. Waibel, "Attention-Passing Models for Robust and Data-Efficient End-to-End Speech Translation," *Transactions of the Association for Computational Linguistics (TACL)*, 2019.

[14] Y. Liu, H. Xiong, J. Zhang, Z. He, H. Wu, H. Wang, and C. Zong, "End-to-End Speech Translation with Knowledge Distillation," in *Proc. Interspeech*, 2019.

[15] M. Stoian, S. Bansal, and S. Goldwater, "Analyzing asr pretraining for low-resource speech-to-text translation," in *ICASSP*, 2020.

[16] S. Bird, "A scalable method for preserving oral literature from small languages," in *The Role of Digital Libraries in a Time of Global Change*, G. Chowdhury, C. Koo, and J. Hunter, Eds. Springer Berlin Heidelberg, 2010.

[17] S. Bird, L. Gawne, K. Gelbart, and I. McAlister, "Collecting bilingual audio in remote indigenous communities," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014.

[18] Ethnologue, "Ethnologue (21st edition)." [Online]. Available: https://www.ethnologue.com/enterprise-faq/how-many-languages-world-are-unwritten-0

[19] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, "An attentional model for speech translation without transcription," in *Proceedings of NAACL-HLT*, 2016.

[20] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[21] Y.-S. Chuang, C.-L. Liu, H.-Y. Lee, and L.-s. Lee, "SpeechBERT: An Audio-and-text Jointly Learned Language Model for End-to-end Spoken Question Answering," *arXiv e-prints*, 2019.

[22] C. Wang, Y. Wu, Y. Du, J. Li, S. Liu, L. Lu, S. Ren, G. Ye, S. Zhao, and M. Zhou, "Semantic mask for transformer based end-to-end speech recognition," in *Interspeech*, 2020.

[23] C. Wang, Y. Wu, S. Liu, M. Zhou, and Z. Yang, "Curriculum pre-training for end-to-end speech translation," in *ACL*, 2020.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[25] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Proc. Interspeech*, 2017, pp. 498–502.

[26] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "Spanbert: Improving pre-training by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2020.

[27] R. Cattoni, M. A. D. Gangi, L. Bentivogli, M. Negri, and M. Turchi, "Must-c: A multilingual corpus for end-to-end speech translation," *Comput. Speech Lang.*, vol. 66, p. 101155, 2021.

[28] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on EMNLP: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018.

[29] H. Inaguma, S. Kiyono, K. Duh, S. Karita, N. E. Y. Soplin, T. Hayashi, and S. Watanabe, "Espnet-st: All-in-one speech translation toolkit," *arXiv preprint arXiv:2004.10234*, 2020.

[30] 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 *Interspeech*, 2019.

[31] P. Bahar, A. Zeyer, R. Schlüter, and H. Ney, "On using specaugment for end-to-end speech translation," 2019.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[33] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[34] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "SpanBERT: Improving pre-training by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, 2020.

[35] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: enhanced language representation with informative entities," 2019.

[36] B. Zhang, I. Titov, B. Haddow, and R. Sennrich, "Adaptive feature selection for end-to-end speech translation," in *Findings of EMNLP*, 2020.

[37] Y. Liu, J. Zhu, J. Zhang, and C. Zong, "Bridging the modality gap for speech-to-text translation," *arXiv preprint arXiv:2010.14920*, 2020.

[38] R. Zheng, J. Chen, M. Ma, and L. Huang, "Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation," *arXiv preprint arXiv:2102.05766*, 2021.