Low-Resource Japanese-English Speech-to-Text Translation Leveraging Speech-Text Unified-model Representation Learning

Tu Dinh Tran, Sakriani Sakti

Japan Advanced Institute of Science and Technology, Japan

{s2110422,ssakti}@jaist.ac.jp

Abstract

Speech-to-text translation (S2TT) has made it critically important to overcome language barriers. Several multilingual datasets have been introduced recently to expand the coverage of multilingual S2TT systems. However, most research works only focus on increasing the number of languages covered. Unfortunately, many of those languages were covered with only a few hours of training data resulting in a low translation performance. This paper proposes utilizing a unified speech-text representation learning framework to overcome the shortage of parallel speech-text datasets in the S2TT system. Although the approach can be utilized for any language pair, we focus on the Japanese-English S2TT task and evaluate it on the publicly available CoVoST 2 dataset. In addition, we also evaluate the S2TT system on our new Japanese-English dataset with sentence ambiguities in which the same spoken utterances can have different translation meanings depending on different prosodic features. We achieve competitive results compared with other state-of-the-art models in CoVoST 2 dataset and provide significant improvement in the more challenging case of our new dataset.

Index Terms: Speech-to-text translation, unified speech-text representation learning, Japanse-English languages, low-resource settings

1. Introduction

Speech translation is an innovative AI technology that offers a solution to break language barriers by mimicking professional interpreters and automatically performing the translation. Speech-to-text translation (S2TT) system is a particular system that translates from source language speech to target language text. Conventionally, the S2TT system consists of automatic speech recognition (ASR) and machine translation (MT) in a cascade manner [1, 2]. Modern S2TT systems perform the direct speech-to-text translation in a single model based on deep learning [3, 4, 5]. Many research groups and companies are progressing, and many speech translation services are now available for several languages but still support fewer than 100 languages. Nearly 7000 living languages that 350 million people speak remain uncovered. Therefore, such technology must be developed, especially for under-resourced languages (UL).

Recently, several projects committed to accelerating the development of technology for UL by leveraging multilingual systems capable of handling multiple languages. As a result, several multilingual datasets have been introduced to expand the coverage of multilingual S2TT systems. MuST-C [6] provides a one-to-many translation dataset that contains 400 hours of speech per language for English to 8 languages. But it only covered major European languages. Europarl-ST [7] is a many-to-many translation dataset that is constructed from the debates held in the European Parliament but also only covered 6 European languages with 30 different translation directions. Facebook and Instagram introduced *No Language Left Behind* (NLLB) [8] for the translation of low-resource languages covering 200 languages but with text data only of about 3000 sentences in each language. Recently, the CoVoST [9] and CoVoST 2 [10] corpora, which are derived from Common Voice [11], support the speech translations from English into 15 languages and from 21 languages to English. However, many of those languages and language pairs were covered with less than 5 hours training data (low resource settings) resulting in a low translation performance. Unfortunately, not many studies have addressed how to deal with this problem.

This paper introduces an end-to-end direct speech translation model to handle low-resource datasets, which only consist of a few hours of speech in training datasets. Inspired by the unified pre-trained spoken language model for both speech and text named SpeechT5 [6], our proposed model learns a unified-modal representation for speech and text in the source language. Then, we leverage the shared representation and parallel text translation dataset to build a bridge between source speech and target text. This method helps overcome the scarcity of our S2TT datasets between different languages. Although the approach can be utilized for any language pair, we focus on the Japanese-English (Ja-En) S2TT task and evaluate it on the CoVoST 2 [4] with only about 3 hours of speech data.

In addition, we also introduce a new dataset for Ja-En S2TT in a low-resource setting. Our dataset provides samples that illustrate how ambiguity in Japanese can affect the quality of translation. It consists of sentence ambiguities in which the same spoken utterances can have different translation meanings depending on different prosodic features. For example, Figure 1 shows how pause could vary the meaning of the sentence "白い 屋根の大きい家" [romaji: "*shiroi* <pause> *yaneno ooki ie*"]. The pause between "*shiroi*" (white) and "*yaneno*" (roof) results in the translation "*A white house with a big roof*" instead of "*A big house with a white roof*". Since speech ambiguation is common in reality, our dataset is essential to train and evaluate speech translation systems for practical circumstances. To our knowledge, this is the first dataset that pays attention to speech ambiguity in Ja-En S2TT. Here, we also evaluate our proposed system on this new dataset compared to the baseline S2TT system.

2. Related works

Numerous works have been focused on improving the performance of low-resource S2TT recently. The direct speech-to-text translation using encoder-decoder model with word-level

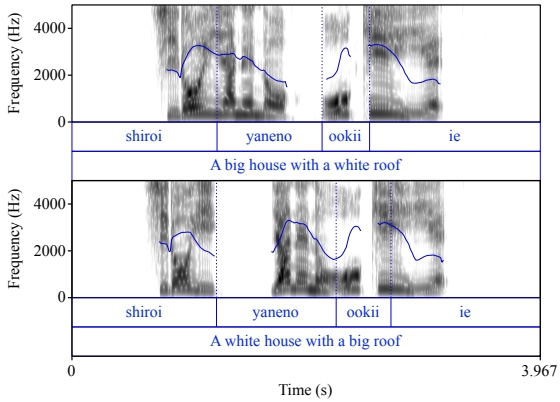

Figure 1: *An example of speech ambiguity.*

decoding is applied in [13]. Multi-source model is fetched with speech and text together to perform low-resource S2TT task in [14]. Other works concentrate on utilizing pre-trained models of related or unrelated languages on ASR task to facilitate low-resource S2TT [15, 16]. To effectively fine-tune acoustic modeling and multilingual text generation model on low-resource S2TT task, only parameters of LayerNorm and Attention modules (LNA) are fine-tuned in [17]. mSLAM [18] focuses on building a multilingual pre-trained model of speech and text, which is trained on 429K hours of unannotated speech data in 51 languages, 15TBs of unannotated text data in 101 languages and especially 2.3k hours of speech-text ASR data with CTC loss to represent both speech and text in a shared representation space. However, for the specific task like S2TT, the direction of source speech to target text is not established well since mSLAM [18] uses the CTC loss only on the ASR task of each language. Additionally, using the encoder-only structure is not ideal for sequence-to-sequence tasks. Our work, which is based on cross-modality SpeechT5 [12] with encoder-decoder structure, enables the model to learn the translations from source speech to target text during the task of text translation before fine-tuning on downstream S2TT task.

3. Proposed method

Obtaining the shared speech-text representations for target side from SpeechT5 [12], our framework then constructs the bridge between source speech and target text through performing text translation task. At last, we fine-tune the model on low-resource S2TT dataset. We cover the overall structure of SpeechT5 [12] in Subsection 3.1 and provide details about our framework in Subsection 3.2.

3.1. SpeechT5

SpeechT5 [12] is a unified model framework that aims to learn contextual representation for both speech and text via a shared encoder-decoder structure. It follows Transformer framework together with six modal-specific pre/post-nets: Speech-encoder Pre-net, Text-encoder Pre-net, Speech-decoder Post-net, Text-decoder Post-net, Speech-decoder Pre-net, Text-decoder Pre-net. The Pre-net modules take raw audio $\mathbf{X}^s = (x_1^t, ..., x_{M^s}^t) \in \mathcal{D}^s$ or text $\mathbf{X}^t = (x_1^t, ..., x_{M^t}^t) \in \mathcal{D}^t$ as input and output the specific vector representations based on modality. The vector representations are fetched through the shared encoder-decoder module as sequence-to-sequence conversion. The post-net modules take responsibility to generate the sequence text or log Mel-filterbank features based on modality.

SpeechT5 adopts the multi-task learning method with three main tasks: text pre-training, speech pre-training, and joint pre-training. The text pre-training task follows BART [19] which trains on text denoising. The corrupted text $\tilde{\mathbf{X}}^t = (\tilde{x}_1^t, ..., \tilde{x}_M^t)$, which are created by using text infilling method as the same as BART [19], are fed into the model to reconstruct the original text $\mathbf{X}^t$. The model is optimized with maximum likelihood estimation as:

$$\mathcal{L}_{mle}^t = \sum_{n=1}^{N^t} \log p(x_n^t | x_{<n}^t, \tilde{\mathbf{X}}^t). \tag{1}$$

Speech pre-training performs on two subtasks: bidirectional masked prediction and sequence-to-sequence generation. Following HuBERT [20], SpeechT5 masks the output $\mathbf{H}$ of Speech-encoder Pre-net and inputs them into the encoder to generate hidden representation $\mathbf{U} = (u_1, ..., u_{N^h})$. The frame-level targets $\mathbf{Z} = (z_1, ..., z_{N^h})$ is generated with the label from the 6-th Transformer layer of the first iteration pre-trained HuBERT base model. Cross Entropy loss of bidirectional masked prediction subtask is formulated as:

$$\mathcal{L}_{mlm}^s = \sum_{n \in \mathcal{M}} \log p(z_n | \tilde{\mathbf{H}}, n), \tag{2}$$

where z_n is frame-level target at timestep n, $\tilde{\mathbf{H}}$ is masked version of $\mathbf{H}$ and $\mathcal{M}$ is set of masked timestep. The subtask sequence-to-sequence requires the model to reassemble the log Mel-filterbank $\mathbf{Y}^f = (y_1^f, ..., y_{N^f}^f)$ from the extracted log Mel-filterbank $\mathbf{X}^f = (x_1^f, ..., x_{N^f}^f)$ of raw audio $\mathbf{X}^s$, given randomly masked input following bidirectional masked prediction. L1 distance is used as the loss function.

$$\mathcal{L}_1^s = \sum_{n=1}^{N^f} \| y_n^f - x_n^f \|_1. \tag{3}$$

In addition, Cross Entropy loss $\mathcal{L}_{bce}^s$ is utilized for training model to predict stop token of the output Mel-filterbank. Joint pre-training leverages paired speech and text for building cross modality-mapping. In specific, the model aligns speech representation and text representation through a shared codebook via the vector quantization method. With a fixed-size codebook $\mathbf{C}^k$ consisting of K learnable embeddings, continuous speech or text representations u_i from the output of the encoder are converted into discrete representations c_i by a quantizer. The L2 distance is utilized as the metric for obtaining the nearest discrete representations c_i for u_i.

$$c_i = \arg\min_{j \in [K]} \| u_i - c_j \|_2, \tag{4}$$

where c_j is the j-th vector in the codebook. In addition, 10% of the contextual text/speech representations is replaced by quantized latent representations, which encourages the model to learn from the quantized latent representations containing the information of both speech and text. Diversity loss is used to spread the attention of the model over different codes by maximizing the entropy of average Softmax distribution.

$$\mathcal{L}_d = \frac{1}{K} \sum_{k=1}^{K} p_k \log p_k, \tag{5}$$

where p_k denotes the average probability of k-th code in the codebook. The total loss function of the model is the sum of the aforementioned losses:

$$\mathcal{L} = \mathcal{L}_{mlm}^s + \mathcal{L}_1^s + \mathcal{L}_{bce}^s + \mathcal{L}_{bce}^s + \mathcal{L}_{mle}^t + \gamma \mathcal{L}_d, \tag{6}$$

where $\gamma = 0.1$ is the coefficient of $\mathcal{L}_d$.

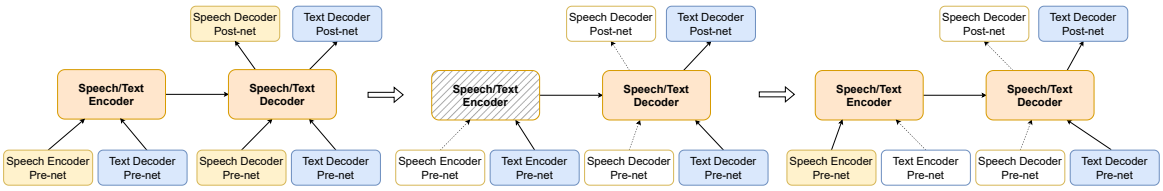

Figure 2: *Our framework based on SpeechT5 [12] contains different modules. In each training step, specific modules are used depending on the training task. Unused modules are denoted by white color and the frozen module is denoted by cross line.*

Table 1: *Examples of our new dataset.*

Japanese sentence	Hint	English translation
食後 に 大人 は 2 錠 、<pause> 子ども は 1 錠です shokugo ni otona wa ni joo komodo wa ichi joodesu	食後 に 大人 shokugo ni otona	Two tablets for adults after meals, and one tablet for children.
食後 に <pause> 大人 は 2 錠、<pause> 子ども は 1 錠です shokugo ni otona wa ni joo komodo wa ichi joodesu	食後 に 2 錠 shokugo ni ni joo	After meals, two tablets for adults and one tablet for children.
ダブル の バスつき の 部屋 が いいです daburu no basutsuki no heya ga iidesu	ダブル の バス daburu no basu	I want a room with double bathrooms.
ダブル の <pause> バスつき の 部屋 が いいです daburu no basutsuki no heya ga iidesu	ダブル の 部屋 daburu no heya	I want a double bedroom with a bathroom.

3.2. Our framework

Our framework (shown in Figure 2) consists of three main steps:

- **Step 1**: pre-train the SpeechT5 model on the monolingual paired dataset of Japanese speech and text.
- **Step 2**: freeze the encoder while fine-tune the pre-trained Japanese SpeechT5 on high resource Ja-En text-to-text translation dataset.
- **Step 3**: fine-tune the model on Ja-En low-resource S2TT dataset.

The first step follows the idea of speechT5 [12] which builds a cross-modality for both Japanese speech and text. In addition, we also obtain a model whose encoder can encode Japanese speech information for the S2TT task. Then, we attain a decoder to decode English text from the contextual information of Japanese text in the second step. Since we obtained cross-modality representations for speech and text from the shared encoder in the first step, freezing the encoder helps the model map from the shared representations to English text. In specific, the translations with the direction from Japanese speech to English text are learned during the second step. At last, we fine-tune the obtained model on the downstream S2TT task. Both text translations and speech translations use Cross Entropy as the loss function. Besides, the high resource of the monolingual dataset of Japanese speech and the bilingual dataset of Ja-En text-to-text translation are key components in our proposed framework.

4. New dataset

The new dataset[1] is manually collected and evaluated with the help of professional translators who are fluent in Japanese and English. Speech ambiguity is related to syntactic ambiguation [21] which means the meaning of a sentence is determined by the different syntactic structures. Prosodic features can be used as cues for syntactic disambiguation [22]. Thus we first collect all Japanese sentences with syntactic ambiguation in books or documents. Each sentence has 2-3 different meanings and we also provide an explanation or hint for each meaning (shown in Table 1). Then, we record the Japanese speech with four different speakers 2 females and 2 males denoted as F01, F02, M01, and M02. The speakers try to make each spoken utterances convey the correct meaning with specific prosodic features like pitch and pause. Audio files are formatted as 16kHz WAV. At last, the translators translate our Japanese sentences into English. Due to the shortage of human resources, our dataset only supports the low-resource settings of S2TT (shown in Table 2). Besides, with four different speakers, the dataset can also support voice conversion or ASR in low-resource settings.

Table 2: *Statistics of our new dataset. #triplets and #hour denote the number of triplets and the number of audio hours, respectively.*

Split	#triplets	#hour
Training	308	0.8
Validation	67	0.13
Test	100	0.2

5. Experimental results

5.1. Implementation details

Our models are implemented following SpeechT5[2] [12] and Fairseq[3] [23]. The encoder-decoder module has 6 encoder layers and 6 decoder layers, where the dimension is 768 and the number of attention heads is 12. The settings of Speech-encoder Pre-net, Text-encoder Pre-net, Speech-decoder Post-net, Text-decoder Post-net, Speech-decoder Pre-net, Text-decoder Pre-net, and codebook follows SpeechT5 [12]. We create a 32K universal BPE vocabulary for Japanese text and another 32K universal BPE vocabulary for English text. Our models are trained

[1] https://github.com/ha3ci-lab/data_stprodis_jaen

[2] https://github.com/microsoft/SpeechT5/tree/main/SpeechT5

[3] https://github.com/facebookresearch/fairseq

Table 3: *BLEU score of baselines and our model on CoVoST 2 and proposed dataset.*

Model	CoVoST 2	Proposed dataset
End-to-End ST [10]	1.5	0.05
LNA [17]	2.1	N/A
mSLAM [18]	3.3	N/A
Cascade ST [10]	**3.8**	N/A
Our model	3.41	**3.23**

Table 4: *Ablation study.*

Model	CoVoST 2	Proposed dataset
Naive SpeechT5	0.84	0.6
Naive SpeechT5 with step 1	3.0	2.79
Naive SpeechT5 with step 2	2.9	2.56
Our model	**3.41**	**3.23**

only on a single A100 GPU.

For pre-training SpeechT5 on Japanese, we combine 66.3 audio hours with transcripts for training from 9 hours of JSUT corpus [24], 52.8 hours of Kokoro[4], 1.3 hours of CoVoST 2 training data and 3.2 hours of proposed dataset training data (audios of 4 speakers). The validation set contains 8.1 audio hours which is also a combination of 1 hour of JUST corpus, 5.8 hours of Kokoro, 0.8 hours of CoVoST 2, and 0.5 hours of our own data. Our settings follow the settings of pre-training step in SpeechT5 except for 80000 updates and 6400 warming-up updates.

In the text translation step, we utilize JESC [25] which consists of 2.7 million sentences for the training set, 2000 sentences for the dev set, and 2000 for the test set. We train our models with 0.001 learning rate, 40000 number of updates, and 4000 number of warming up updates. The best model is chosen based on SacreBLEU [26] with beam size 5 over the validation set.

To evaluate our models on Ja-En low-resource S2TT dataset, we use our new dataset recorded by speaker F01 and CoVoST 2 dataset which has 1.3 hours for the training set, 0.8 hours for the validation set, and 0.8 hours for the test set. We use the same method as text translation to select the best model for testing.

5.2. Results and discussion

We compare our models with multiple baselines. End-to-End ST [10] is the direct speech-to-text translation model with Transformer framework where the encoder is pre-trained with ASR using English data in CoVoST 2. LNA [17] is also a direct speech translation model which is initialized by Wav2vec 2.0 [27] for the encoder and mBART 50 [28] for the decoder. mSLAM [18] adopts the Conformer framework with 2 billion parameters and is pre-trained on a multilingual dataset of unlabeled text and speech as well as paired speech-text. Although we propose a direct speech translation model, we also compare with Cascade ST [10] which is composed of a large Transformer pre-trained ASR model and multilingual text translation trained on all X-En and En-X.

In Table 3, the results show that our model outperforms other direct speech-to-text translation models on CoVoST 2 with 3.41 BLEU score. This indicates the effectiveness of our proposed method which indirectly trains the model on the direction of Japanese speech to English text. Besides, we also show the effective way to utilize available resources in Japanese. Although our model cannot beat Cascade ST, it still narrows the gap between the direct translation and the traditional cascade solution. In addition, our model achieves a decent recent result of 3.23 BLEU score on the proposed dataset despite the scarcity of the dataset and the speech ambiguity.

To further investigate the effect of different steps in our framework, we also conduct the ablation study. The results are shown in Table 4. First, we directly train the naive SpeechT5 model on each speech translation dataset without pre-training on the paired dataset of Japanese speech and text or fine-tuning on the text translation dataset. This model shows poor results since it suffers from the scarcity of training data without knowledge from the preceding steps. Second, pre-training the model for cross-modal representations before training the downstream task shows improvement which comes from the pre-trained Japanese speech encoder. Third, the second step also improves the performance of naive SpeechT5 since it provides the knowledge about target side. Fourth, our full model achieves the best result since pre-training on the two preceding steps enables the model to learn the translation from source speech to target text.

6. Conclusion

In this paper, we propose an efficient framework for low-resource S2TT. We exploit SpeechT5 [12] to build a shared representation for both Japanese speech and text. Then, we fine-tune the model on high-resource Ja-En text translation, which helps train Japanese speech to English text thanks to the pre-trained cross-modality encoder. At last, we fine-tune the obtained model from the preceding steps for low-resource S2TT. We also compare the performance of our models with other those of state-of-the-art methods on CoVoST 2 and the proposed dataset. We outperform the end-to-end direct translation baselines and close the gap with the conventional cascade approach. Besides, we also introduce a new low-resource S2TT dataset that focuses on speech ambiguity with different prosodic features and achieve decent results with our model.

In this study, we only investigated Japanese-English speech translation, but the framework can generally be applied to other low-resourced language pairs. In the future, we will apply this method to other languages. In addition, we will explore extracting and incorporating prosodic features, especially pitch and pause to better distinguish speech ambiguity in S2TT.

7. Acknowledgements

Part of this work is supported by JSPS KAKENHI Grant Numbers JP21H05054 and JP21H03467.

[4] https://github.com/kaiidams/Kokoro-Speech-Dataset

8. References

[1] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, "The atr multilingual speech-to-speech translation system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 365 – 376, 04 2006.

[2] S. Sakti, M. Paul, A. Finch, S. Sakai, T. T. Vu, N. Kimura, C. Hori, E. Sumita, S. Nakamura, J. Park, C. Wutiwiwatchai, B. Xu, H. Riza, K. Arora, C. M. Luong, and H. Li, "A-star: Toward translating asian spoken languages," *Computer Speech & Language*, vol. 27, no. 2, pp. 509–527, 2013, special Issue on Speech-speech translation.

[3] 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," *ArXiv*, vol. abs/1904.06037, 2019.

[4] T. Kano, S. Sakti, and S. Nakamura, "End-to-end speech translation with transcoding by multi-task learning for distant language pairs," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1342–1355, 2020.

[5] A. Lee, H. Gong, P.-A. Duquenne, H. Schwenk, P.-J. Chen, C. Wang, S. Popuri, Y. Adi, J. Pino, J. Gu, and W.-N. Hsu, "Textless speech-to-speech translation on real data," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States, 2022, pp. 860–872.

[6] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: a Multilingual Speech Translation Corpus," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 2012–2017.

[7] J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, J. Jorge, N. Roselló, A. Giménez, A. Sanchis, J. Civera, and A. Juan, "Europarl-st: A multilingual corpus for speech translation of parliamentary debates," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8229–8233.

[8] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard *et al.*, "No language left behind: Scaling human-centered machine translation," *arXiv preprint arXiv:2207.04672*, 2022.

[9] C. Wang, J. Pino, A. Wu, and J. Gu, "CoVoST: A diverse multilingual speech-to-text translation corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, May 2020, pp. 4197–4203.

[10] C. Wang, A. Wu, J. Gu, and J. Pino, "CoVoST 2 and Massively Multilingual Speech Translation," in *Proc. Interspeech 2021*, 2021, pp. 2247–2251.

[11] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, 2020, pp. 4218–4222.

[12] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li, and F. Wei, "SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, May 2022, pp. 5723–5738.

[13] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Low-Resource Speech-to-Text Translation," in *Proc. Interspeech 2018*, 2018, pp. 1298–1302.

[14] A. Anastasopoulos and D. Chiang, "Leveraging Translations for Speech Transcription in Low-resource Settings," in *Proc. Interspeech 2018*, 2018, pp. 1279–1283.

[15] M. Stoian, S. Bansal, and S. Goldwater, "Analyzing asr pretraining for low-resource speech-to-text translation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7909–7913, 2019.

[16] 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 the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jun. 2019, pp. 58–68.

[17] X. Li, C. Wang, Y. Tang, C. Tran, Y. Tang, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Multilingual speech translation from efficient finetuning of pretrained models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 827–838.

[18] A. Bapna, C. Cherry, Y. Zhang, Y. Jia, M. Johnson, Y. Cheng, S. Khanuja, J. Riesa, and A. Conneau, "mslam: Massively multilingual joint pre-training for speech and text," *arXiv preprint arXiv:2202.01374*, 2022.

[19] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.

[20] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *Proc. IEEE/ACM Trans. Audio, Speech and Lang.*, vol. 29, p. 3451–3460, oct 2021.

[21] K. SUZUE and Y. MASATAKA, "Structural ambiguity resolution in the process of reanalysis: Evidence from japanese sentence comprehension," *Tohoku Psychologica Folia*, vol. 76, pp. 20–37, 2017.

[22] Y. Misono, R. Mazuka, T. Kondo, and S. Kiritani, "Effects and limitations of prosodic and semantic biases on syntactic disambiguation," *Journal of Psycholinguistic Research*, vol. 26, pp. 229–245, 1997.

[23] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019, pp. 48–53.

[24] R. Sonobe, S. Takamichi, and H. Saruwatari, "Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis," *arXiv preprint arXiv:1711.00354*, 2017.

[25] R. Pryzant, Y. Chung, D. Jurafsky, and D. Britz, "JESC: Japanese-English Subtitle Corpus," *Language Resources and Evaluation Conference (LREC)*, 2018.

[26] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 186–191.

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

[28] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual translation with extensible multilingual pretraining and finetuning," *arXiv preprint arXiv:2008.00401*, 2020.