KHMER SPEECH TRANSLATION CORPUS OF THE EXTRAORDINARY CHAMBERS IN THE COURTS OF CAMBODIA (ECCC)

Kak Soky[⋆†]*, Masato Mimura*[⋆]*, Tatsuya Kawahara*[⋆]*,*
Sheng Li[†]*, Chenchen Ding*[†]*, Chenhui Chu*[⋆]*, Sethserey Sam*[‡]

[⋆]Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto, Japan
[†]National Institute of Information and Communications Technology (NICT), Kyoto, Japan
[‡]Cambodia Academy of Digital Technology (CADT), Phnom Penh, Cambodia

ABSTRACT

Speech translation (ST) is a subject of rapidly increasing interest in the area of speech processing research. This interest is apparent from the increasing tools and corpora for this task. However, the lack of sufficient datasets is still the biggest challenge for under-resourced languages. Specifically, ST requires a large corpus of parallel speech, transcription, and translation text. In this work, we construct a large corpus of the Extraordinary Chambers in the Courts of Cambodia (ECCC), including simultaneous translation from Khmer into English and French. We also address the problem of sentence segmentation of Khmer by conducting a bilingual sentence alignment from English to Khmer with a monotonic assumption. This corpus has approximately 155 hours of speech in length and 1.7M words of text. We also report the baseline results of automatic speech recognition (ASR), machine translation, and ST systems, which show reasonable performance.

Index Terms— Khmer language, low-resource language, spoken language translation corpus, court dataset

1. INTRODUCTION

In the last decade, the advancement of deep learning techniques and computing resources has been successfully boosting end-to-end (E2E) models [1, 2, 3, 4, 5] to achieve promising results in various systems. For instance, E2E speech translation (ST) is a system that directly translates the speech signals in a language to the text of another language. It integrates automatic speech recognition (ASR) and machine translation (MT) which are used in the traditional approach of the cascading models into a single model. However, E2E-ST requires the parallel resources of source-language speech and target text in another language, which is currently available for a limited number of language pairs and in a limited amount.

There are several spoken-language translation (SLT) corpora that are available in a single speech source language, such as Must-C [6], Fisher-CallHome Spanish [7], and in multiple speech source languages, including Europarl-ST [8] and Multilingual TEDx [9]. Among them, only Europarl-ST is simultaneous ST. However, it has less than 50 hours in non-English source speech and less than 90 hours in English source speech. In this work, we build a large Khmer SLT corpus by collecting approximately 200 hours of raw audio of the Extraordinary Chambers in the Courts of Cambodia (ECCC) in Khmer and the corresponding documents in three languages: Khmer, English, and French.

Sentence alignment of the source and target language is a crucial stage in SLT corpus creation. Better language processing tools are required to make better quality alignment and for time efficiency. However, this assumption does not hold for most low-resource languages, which usually have worse performance or lack of toolkit to support those languages. Additionally, the written style of Khmer occasionally uses spaces only to make the text more natural for reading; sometimes, there are no sentence boundaries or punctuation marks to separate the text sentences. To overcome these challenge characteristics, we propose aligning the bilingual sentences in a monotonic process that only requires the sentence segmentation of the source-language text, whereas only word tokenization is required for the target-language text. This is suitable for a simultaneous translation dataset such as the ECCC.

Another potential challenge is text-to-speech alignment. Most other corpora have timestamp information for the audio data, whereas it is not available for the original ECCC dataset. Therefore, we generated timestamps for the speech data that corresponded to each sentence of the text. Ultimately, we created a large Khmer SLT corpus of the ECCC, which has more than 150 hours in length of speech in Khmer, approximately 65K utterances in each language pair of Khmer-English and Khmer-French. In this corpus, 60% of speech is the original speech of Khmer speakers, and 40% of speech is translated from English and French. Moreover, there is a wide range of speakers: witnesses, defendant, lawyers, judges, and officers.

The rest of this paper is structured as follows. Section 2 gives a brief description of the ECCC dataset. We present the detailed methods of the corpus creation in Section 3. In Section 4, we describe the setup and results of the experiments

978-1-6654-0870-7/21/$31.00

for ASR, MT, and ST. We conclude the paper in Section 5.

2. ECCC BACKGROUND

The ECCC is a court established to prosecute the senior leaders who committed crimes during the Khmer Rouge regime in Cambodia from 1975 to 1979, a regime known as Democratic Kampuchea. The trials have been subsequently divided into four cases that began on February 17, 2009. However, as of this work, these trials are still in progress, and only a tiny part of the resources have been released to the public. Therefore, we chose only the first case, which spanned from February 17 to November 27, 2009, as the resources of that case are available.

The trial had two kinds of hearing: public and non-public. Each hearing was simultaneously conducted in three languages: Khmer, English, and French. This means that the videos were recorded in the courtroom in the language of the main speaker. Concurrently, the human translators translated that speech to the other two languages. Each video, therefore, has three different languages. Thus far, the recordings have been carefully transcribed by native transcribers. Each transcription has the transcription of a single day of the trial, which corresponds to four or five audio sessions. Each recording session has a length of 5 to 150 minutes and involves a wide range of speakers: witnesses, the defendant, judges, clerks or officers, co-prosecutors, experts, defense counsels, civil parties, and interpreters. As a result, we have collected 222 recording sessions that correspond to 60 documents. Each transcription has many pages of A4 size, ranging from 5 to 200. Finally, the public hearing videos are uploaded to a YouTube channel[1], and the proceedings are published in a digital format at the ECCC official website[2].

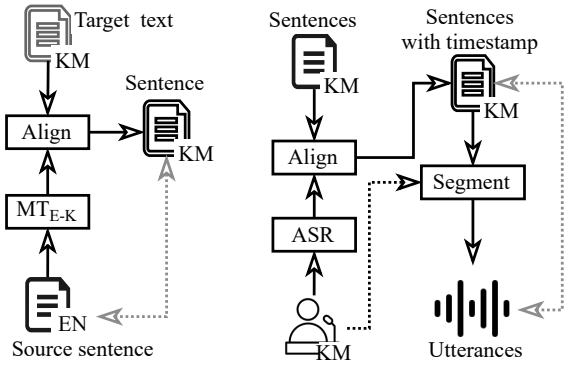
In [10], there is a bilingual Khmer-English ECCC corpus for MT, which has only text data. In that work, an evaluation was not conducted, although the corpus size of the dataset was reported. In this work, our main target is to construct a speech corpus, so that ASR of Khmer and ST of Khmer to English and French can be conducted. We use a subset of the raw text corpus in [10]; however, we conducted the data cleansing and sentence alignment with different methods.

3. CORPUS CONSTRUCTION

The presented raw resources in Section 2 are useful for ST, ASR, and MT systems. However, it is not possible to directly use them for those tasks. Furthermore, this dataset lacks timestamps. We considered sentence alignment as a critical component of corpus creation. English has better language processing tools: consequently, we used it as the source language for the alignment purpose because of this abundance of supporting language tools.

[1]https://www.youtube.com/user/krtribunal/

[2]https://www.eccc.gov.kh/



(a) Sentence alignment (b) Text-speech alignment

Fig. 1: The Process of creating the ECCC corpus: (a) bilingual sentence alignment, (b) text-to-speech alignment and segmentation

3.1. Source to Target Sentence Alignment

To align sentences, sentence segmentation is required in both source and target text, as presented in [11, 12, 13]. In these works, the sentences were aligned on the basis of the alignment score of each sentence. With this scoring, the alignment can be in the form of zero-to-one, one-to-zero, one-to-one, one-to-many, many-to-one, and many-to-many. However, only one-to-one is usable in the translation task. Thus, many of the original resources can be removed. Some languages such as Khmer, though, do not have any sentence tokenization tools that support them, not even Moses [14] or Punkt [15]. However, the simultaneous translation is processed in a monotonic and continuous alignment. With this characteristic, only the source language requires sentence segmentation.

We followed Fig. 1a to align the bilingual source and target texts. We first conducted sentence segmentation of English using Moses. The sentences were re-split on the basis of some conjunction words to ensure that they were less than 200 characters (without spaces). We then translated those sentences to the target languages, Khmer and French, using the translation API in Google Sheets. For the ground-truth of Khmer and French, we merged all text into a single line. However, the Khmer language is written without word boundaries. Thus, the Khmer word segmentation tool [16] was used to segment both the translated and ground-truth text.

Second, the alignment between translated and ground truth was conducted using dynamic programming (DP) in a monotonic manner. Sentence boundary tokens were inserted in accordance with the sentence boundaries of the translated text. In this alignment, the calculation was based on word-level Levenshtein distance. As a result, only one-to-zero and one-to-one alignments are obtained. At this point, we removed the one-to-zero aligned sentences from the source language. As result, we obtained $82,078$ sentences in En-

Table 1: Statistics of target languages of the ECCC Khmer SLT corpus

Target	#utterance	#target word	#vocabulary	#hour (train/dev/test)	#avg. target word	#avg. length (s)
EN	65,391	1.24M	15K	139/8/8	19	8.5
FR	64,203	1.33M	21K	136/8/8	21	

Table 2: Statistics of Khmer source text

Language	#word	#vocabulary	#avg. word
KM	1.66M	9K	25

glish aligned with 78,981 sentences in French and 80,417 sentences in Khmer, which means that only 4% and 2% in French and Khmer was discarded, respectively.

3.2. Text to Speech Alignment

Fig. 1b shows the process of the text to speech alignment. We first trained a new acoustic model that supported Vosk[3] using the Basic Expressions Travel Corpus [17] that was used in [18]. Vosk enables us to diarize the speech to generate the transcription with its corresponding timestamp.

Then, we conducted sentence alignment between the segmented sentences and the pseudo labels of ASR diarization output. At this stage, the alignment algorithm in subsection 3.1 was used to generate the ground-truth sentence with the corresponding timestamp. Each sentence has the starting and ending timestamps, which is aligned with a short segment of the audio data.

3.3. Text Cleansing

To cleanse the text corpus, we focused on the transcribed text that corresponds to speech data using the following process: removing unrelated parts (no speech) such as page headings, descriptions of the activity, or feelings that are usually marked by "[ ]". For English and French, the text normalization was conducted using Moses. Subsequently, the punctuation marks were removed and the text was made lowercase. For Khmer, we deleted the non-standard characters, punctuation marks, and other Latin symbols. We also normalized the text to correct the spelling and order of the Khmer characters or diacritics, as presented in [19]. The numbers and abbreviations were also replaced by their standard spoken equivalent in all languages.

3.4. Speech Cleansing

In this cleansing process, we had to ensure that the length of each audio segment was usable in ASR. A usable length is in range from $3s$ to $30s$. Each sentence of the transcription or translation text had to be less than 300 characters in length because each source sentence in English was limited to less than 200 characters before alignment. Sentences and audio segments that did not meet these criteria were deleted from the corpus. After the cleansing process, 19% of the original resources were discarded for Khmer to English and 20% for Khmer to French. Many resources were discarded because the performance of the Khmer ASR model was insufficient for transcribing some parts of the speech in this dataset. This is related to the domain and speaking-style mismatch.

[3]https://alphacephei.com/vosk/

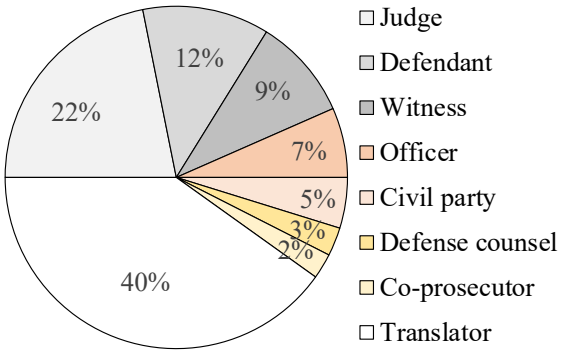


Fig. 2: Speaker distribution in the ECCC Khmer SLT corpus

3.5. Corpus Statistics

Table 1 gives the statistics of Khmer-source speech translation to English and French. In terms of speech, the average length of each utterance is 8.5s. In total, it is approximately 155 hours of speech in the translations to English and to French, which is about 65K utterances. In terms of text, the target sentence in French and English have respectively 19 and 21 words on average, whereas it has 25 words on average in the Khmer source sentence, as presented in Table 2. Overall, the text size is approximately 1.6M words and the vocabularies are 9K, 15K and 21K in Khmer, English and French, respectively. Finally, each language pair was split into a training, development and test sets.

The graph in Fig. 2 shows the speaker distribution for each speaker group. It is shown that 40% of the speech is that of translators who translated the speech from native English and French speakers. These speakers include judges, officers, co-prosecutors, defense counsels, civil parties, and experts. The remaining 60% of speech is the original speech of Khmer speakers. The largest percentage is a speech of the judges, which makes up 22% of the corpus, followed by 12% from

Table 3: Word error rate (WER) of the ASR models in Khmer; '*' model is used in cascade-ST and E2E-ST

Transformer ASR Model	WER
w/o augmentation	23.6
w/ speed perturbation (SP)	22.2
w/ SpecAugment (SA)	21.8
w/ SP + SA *	**21.4**

Table 4: BLEU for translation from Khmer

KM→Target	BLEU		
	MT	Cascade-ST	E2E-ST
EN	16.63	15.14	13.81
FR	11.53	10.66	9.39

the defendant, 9% from the witnesses, and 17% in total from other speakers.

4. EXPERIMENTS

To evaluate this corpus, we conducted ASR, MT, and ST experiments using a Transformer-based [4] architecture implemented in ESPnet [20]. In all experiments, the network is comprised of six encoder layers and six decoder layers. The dimension of feed-forward network was set to 2,048, and dropout was set to 0.1. The model used 4-head self-attention of 256 dimensions. We trained each model on a single 12-GB GPU Titan X (Pascal) with the aforementioned configurations.

4.1. Automatic Speech Recognition (ASR)

In the ASR system, we trained the model using 80-dimensional log-melscale filterbank (lmfb) coefficients and 3-dimensional pitch features. This network was started on downsampling by a 2-layer time-axis convolutional layer with 256 channels, stride size 2, and kernel size 3. The model was jointly trained with connectionist temporal classification (CTC) (weight $\alpha = 0.3$) for 45 epochs with a batch size of 64. The Noam optimizer was used with 25K warmup steps and an initial learning rate of 5.

The transcription was stripped of all punctuation marks. We used 5K byte-pair encoding (BPE) tokens [21] as the vocabularies for each language. Speech perturbation [22] and SpecAugment [23] were applied as the speech data augmentation. All the system performances are evaluated in WER and shown in Table 3.

4.2. Machine Translation (MT)

For MT, we trained another Transformer-based model for 100 epochs with a batch size of 96. The Noam optimizer was used with 8K warmup steps and an initial learning rate of 1. In each language pair, all punctuation marks were stripped and converted to lowercase for English and French. We applied 10K BPE tokens of the combination of source and target vocabularies, which resulted in 5K per language. The translation performances are reported using BLEU, as shown in Table 4.

4.3. Speech Translation (ST)

The ST front-end configuration is similar to the ASR system. The speed perturbation and SpecAugment were applied as the speech data augmentation. The 10K BPE tokens of joint source and target vocabularies were used as they were for MT. However, the training was conducted for 60 epochs with a batch size of 64. The ASR and MT pre-trained models, which were presented in the previous Sections 4.1 and 4.2, were respectively used to initialize the E2E-ST encoder and decoder. With this initialization, the E2E-ST can achieve a reasonable performance, as described in [24]. For cascade-ST, the output of the ASR system was translated into the target languages using the MT models. The results are reported in Table 4.

4.4. Discussion

Table 4 shows the translation from Khmer to English performs better than Khmer to French for all translation systems. This is reasonable because English was directly used as the source language in the bilingual sentence alignment to Khmer and French, whereas Khmer to French was indirectly aligned. Moreover, these results show that the E2E-ST performs worse than the cascade-ST because of the non-monotonic alignments of speech-text.

In Table 5, the ASR performance for each speaker group is presented. The WER of the witnesses was higher than that of the other speakers. The main reason for this is that the witnesses are the victims of the Khmer Rouge regime, and most of them are illiterate in the Khmer language. They sometimes cannot pronounce words correctly and they exhibit disfluency and emotions in their speech during the trial, compared to the other groups.

Similarly, the translation quality of the witnesses is also worse than that for other groups of speakers, whereas the judges were better in all of the translation models, including MT, cascade-ST, and E2E-ST, which indicates that usually, the judges are well-prepared for their speech.

5. CONCLUSION

In this work, we created the largest ever SLT corpus of Khmer language, including the simultaneous speech of the translators. We extracted the SLT of Khmer to English and French from an ECCC dataset of 222 sessions to build the 155 hours in length of speech and 1.7M words in text. Furthermore, we conducted E2E ASR, MT and ST experiments on the constructed corpus and obtained reasonable performance.

Table 5: Evaluation of system performance for each speaker group

Speaker (#)	#utterance	#word	#hour	ASR	MT (BLEU)		Cascade-ST (BLEU)		E2E-ST (BLEU)	
				WER	EN	FR	EN	FR	EN	FR
Witness (5)	2,068	44.6K	5	23.4	15.63	10.51	13.74	9.84	12.14	8.30
Co-prosecutor (2)	796	22.2K	2	19.7	18.85	12.31	17.15	11.28	16.56	10.77
Civil party (1)	246	7K	0.7	15.3	17.77	12.47	16.07	11.68	14.35	10.10
Judge (2)	92	2.4K	0.3	17.0	21.88	18.71	20.55	15.67	19.27	14.56

We have kept a large proportion of the original corpus by using monotonic sentence alignment and word-based distance calculation. This alignment requires segmentation only the sentences in the source language. This is very effective and useful for aligning from a rich-resource language to other low-resource languages. Moreover, this alignment method will be helpful for applications to similar datasets such as meetings, classroom lectures, and TV programs.

6. REFERENCES

[1] Alex Graves, Santiago Fernandez, Faustino Gomez, and Jurgen Shmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[2] Alex Graves, "Sequence Transduction with Recurrent Neural Networks," *ICML Representation Learning Workshop*, Nov. 2012.

[3] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[4] Ashish Vaswani, Noam Shazeer abd Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[5] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020.

[6] Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco 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, (NAACL-HLT 2019)*, 2019.

[7] Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur, "Improved speech-to-text translation with the Fisher and Callhome Spanish–English speech translation corpus," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, 2013.

[8] Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan, "Europarl-st: A multilingual corpus for speech translation of parliamentary debates," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[9] Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, R. Cattoni, M. Negri, M. Turchi, D. Oard, and Matt Post, "The multilingual tedx corpus for speech recognition and translation," *ArXiv*, vol. abs/2102.01757, 2021.

[10] Toshiaki Nakazawa, Nobushige Doi, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Yusuke Oda, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi, "Overview of the 6th workshop on Asian translation," in *Proceedings of the 6th Workshop on Asian Translation*, Hong Kong, China, Nov. 2019, pp. 1–35, Association for Computational Linguistics.

[11] Fabienne Braune and Alexander Fraser, "Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora," in *Coling 2010: Posters*, 2010.

[12] Chris Dyer, Victor Chahuneau, and Noah A. Smith, "A simple, fast, and effective reparameterization of IBM model 2," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013.

[13] Brian Thompson and Philipp Koehn, "Vecalign: Improved sentence alignment in linear time and space," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[14] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 2007.

[15] Tibor Kiss and Jan Strunk, "Unsupervised multilingual sentence boundary detection," *Comput. Linguist.*, vol. 32, no. 4, pp. 485–525, Dec. 2006.

[16] V. Chea, Y. Kyaw Thu, C. Ding, M. Utiyama, A. Finch, and E. Sumita, "Khmer word segmentation using conditional random fields," in *Proc. Khmer Natural Language Processing (KNLP)*, 2015.

[17] Gen-ichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto, "Creating corpora for speech-to-speech translation," in *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003, Geneva, Switzerland*, 2003.

[18] Kak Soky, Sheng Li, Tatsuya Kawahara, and Sopheap Seng, "Multi-lingual transformer training for khmer automatic speech recognition," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019.

[19] Benjamin Marie, Hour Kaing, Aye Myat Mon, Chenchen Ding, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita, "Supervised and unsupervised machine translation for Myanmar-English and Khmer-English," in *Proceedings of the 6th Workshop on Asian Translation*, 2019.

[20] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. Interspeech 2018*, 2018.

[21] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016.

[22] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio Augmentation for Speech Recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[23] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019.

[24] Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji 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.