International Journal of Asian Language Processing
© World Scientific Publishing Company

TriECCC: Trilingual Corpus of the Extraordinary Chambers in the Courts of Cambodia for Speech Recognition and Translation Studies

Kak Soky, Masato Mimura, Tatsuya Kawahara, Chenhui Chu
Graduate School of Informatics, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan
{soky,mimura,kawahara}@sap.ist.i.kyoto-u.ac.jp

Sheng Li, Chenchen Ding
National Institute of Information and Communications Technology,
Soraku-gun, Kyoto 619-0289, Japan
{sheng.li,chenchen.ding}@nict.go.jp

Sethserey Sam
Cambodia Academy of Digital Technology (CADT),
Phnom Penh 12252, Cambodia
sethserey.sam@cadt.edu.kh

Received (Day Month Year)
Revised (Day Month Year)

This article presents an extended work on the trilingual spoken language translation corpus of the Extraordinary Chambers in the Courts of Cambodia (ECCC), namely TriECCC. TriECCC is a simultaneously spoken language translation corpus with parallel resources of speech and text in three languages: Khmer, English, and French. This corpus has approximately 62 thousand utterances, approximately 146, 148, and 125 hours in length of speech, and 1.6, 1.2 and 1.3 million words in text, in Khmer, English, and French, respectively. We first report the baseline results of automatic speech recognition, machine translation (MT), and speech translation (ST) systems, which show reasonable performance. We then investigate the use of the ROVER method to combine multiple MT outputs and fine-tune the pre-trained English-French MT models to enhance the Khmer MT systems. Experimental results show that the ROVER is effective for combining English-to-Khmer and French-to-Khmer systems. Fine-tuning from both single and multiple parents shows the effective improvement on the BLEU scores for Khmer-to-English/French and English/French-to-Khmer MT systems.

Keywords: Khmer language; low-resource language; trilingual corpus; court speech; automatic speech recognition, machine translation, spoken language translation

1. Introduction

In the last decade, 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 applications of speech and language processing. For instance, E2E speech translation (ST) directly translates the speech signals in one language to the text of another language. It integrates automatic speech recognition (ASR) and machine translation (MT) 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. In this work, we target to build a large parallel trilingual spoken language translation (SLT) corpus of the Extraordinary Chambers in the Courts of Cambodia (ECCC) in Khmer, English, and French, namely TriECCC, which has both speech and text for each language.

There are several SLT corpora 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 extract the parallel speech of approximately 200 hours of raw audio from ECCC and its corresponding documents in Khmer, English, and French. This corpus will be not only usable for a pure ASR, MT, and ST, but also for a wide range of advanced tasks including multilingual ASR, MT, ST, cross-lingual, multi-source translation[10,11,12,13,14,15,16], or joint training[17,18].[a]

Sentence alignment of the source and target language is crucial in SLT corpus creation. Better language processing tools are required to improve quality alignment and 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; however, 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. In contrast, only word tokenization is needed for the target-language text. This is suitable for a simultaneous translation dataset like the ECCC or Europarl. Secondly, we apply the Recognizer Output Voting Error Reduction (ROVER) method[19], a voting mechanism of multiple automatic speech recognition outputs, to improve the quality of the bilingual sentence alignment to Khmer by voting the alignment outputs of English-Khmer and French-Khmer.

Another challenge is text-to-speech alignment. Most other corpora have timestamp information for the audio data, but it is unavailable 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 parallel TriECCC, which respectively has about 146, 148, and 125 hours in length of speech in Khmer, English, and French, approximately 62K utterances in each language pair of six

[a] The data copyright belongs to NICT, Kyoto Univ. speech lab. and CADT, formerly NIPTICT.

directions. In this corpus, 60% of speech is the original speech of Khmer speakers, 18% of speech is the original speech of English speakers, and 22% of speech is the original speech of French speakers. Moreover, there is a wide range of speakers, including witnesses, defendants, lawyers, judges, and officers.

Following our previous work[20], we first evaluate the baseline model of ASR, MT, and both cascaded-ST and E2E-ST on Transformer-base architecture[4] using the TriECCC. Among them, Khmer language systems show worse performance than other language pairs. In this work, we focus on improving the Khmer MT from/to English and French. We first investigate the system combination of using the ROVER method for combining MT outputs. We then fine-tune the MTs of Khmer language pairs using the pre-trained models of English-French MTs to initialize encoder or decoder of each Khmer MT model. Experimental results show that the fine-tuning process improves the BLEU scores on Khmer-to-French, French-to-Khmer, Khmer-to-English, and English-to-Khmer MT systems.

The main contributions of this work compared to our previous work[20] are as follow:

- We extend a single Khmer source speech to a parallel simultaneous SLT corpus in three source speeches of Khmer, English, and French, which has approximately 150 hours of speech in each source of six translation directions.
- We extend the baseline E2E systems evaluations of ASR, MT, ST, and cascade-ST for Khmer by adding English and French as the source language using this TriECCC.
- We investigate the use of the ROVER system combination to align sentences from English- and French-to-Khmer and to combine multiple MTs for improvement.
- We evaluate the effective use of a rich-resource pre-trained MT model to enhance the performance of low-resource language Khmer MT system in both single and multiple parents fine-tuning approaches. This method is practical for improving MT performance and reducing training time.

2. TriECCC

2.1. *Khmer language*

Khmer or Cambodian is the official language of Cambodia. Around 90% of Cambodian populations speak this language in Cambodia, and some speakers live in other countries. Khmer language (Cambodian) is one of the under-resourced Southeast Asian languages for natural language processing (NLP). It has an SVO (Subject, Verb, and Object) syntax structure. Syntactically it is pretty similar to Chinese and English, and also it is similar to Japanese, Chinese and Myanmar in the word composition. Each Khmer word is composed of single or multiple syllables, usually not separated by white spaces. Although spaces are used for separating phrases for easy reading, it is not strictly necessary. In addition, these spaces are rarely used in short sentences, and there is no exact rule how they are used. There are three main word groups in modern Khmer: (1) original Khmer words, (2) Sanskrit and Pali, which have been influenced by the royal and religious registers, through Hinduism and Buddhism, and (3) loanwords from French and English, i.e., many words were borrowed and have become a part of the colloquial language, as well as medical and technical terms. There is also a smattering of Chinese and neighboring countries' loanwords in colloquial speech. Unlike Thai, Vietnamese and Lao, Khmer is non-tonal. And it has a high percentage of disyllabic words which are derived from monosyllabic bases by prefixation and suffixation[21].

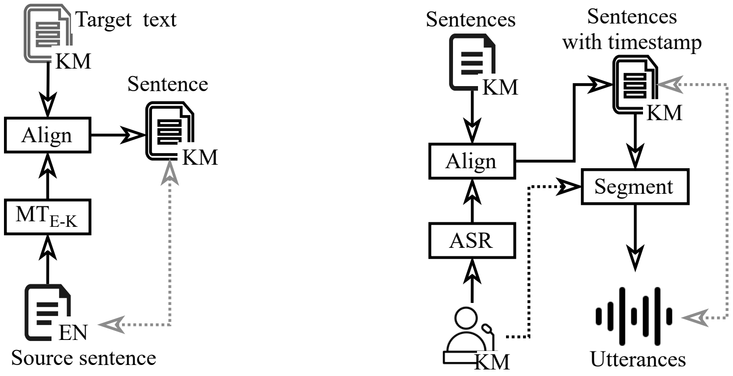
2.2. *ECCC background*

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, known as Democratic Kampuchea. The trials have been subsequently divided into four cases that began on February 17, 2009. These trials are still in progress, and only a small part has 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 covers 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, defendants, 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 in A4 size, ranging from 5 to 200. Finally, the public hearing videos are uploaded to a YouTube channel[b], and the proceedings are published in a digital format at the ECCC's official website[c].

The ECCC dataset has been built as a bilingual Khmer-English corpus for MT, which has only text data[22], a Khmer speech-to-text corpus for ASR[23], and an SLT corpus of the Khmer to English and French by our team[20]. In this work, we extend our previous work[20] by building a trilingual SLT corpus of Khmer, English, and French, which has six SLT directions.

[b] https://www.youtube.com/user/krtribunal/
[c] 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

2.3. *Corpus creation and key statistics*

The raw resources presented in Section 2.2 are useful for ASR, MT, and ST systems. However, it is not possible to directly use them for those tasks, particularly because this dataset lacks timestamps. We considered sentence alignment as a critical component of corpus creation. As English has better language processing tools, we used it as the source language for the alignment purpose.

2.3.1. *Source to target sentence alignment*

To align sentences, sentence segmentation is required in both source and target text, as presented in[24,25,26]. In these works, the sentences were aligned based on 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, however, do not have any sentence tokenization tools such as Moses[27] and Punkt[28]. On the other hand, 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 based on some conjunction words to ensure 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[29] was used to segment both the translated and ground-truth text.

Table 1: CER in source to target sentence alignment

Sentence alignment	**CER (%)**
Bilingual English (EN)-to-Khmer (KM)	13.2
ROVER ({EN, French (FR)}-to-KM)	**12.7**

Table 2: Source to target sentence alignment examples

Reference	1. interrogators and the cadres from Prey Sar would be called to attend such a political session in general 2. but there was another political session conducted separately. 1. រួម ទាំង កង សួរ ចម្លើយ រួម ទាំង ខាង ព្រៃ ស គឺ ជា វគ្គ នយោបាយ រួម ទូទៅ ។ 2. តែ ចំពោះ អ្នក សួរ ចម្លើយ គឺ នយោបាយ ខុស គ្នា នេះ ជា នយោបាយ ជំនាន់ នោះ ។
Monolingual	1. រួម ទាំង កង សួរ ចម្លើយ រួម ទាំង ខាង ព្រៃ ស គឺ ជា វគ្គ នយោបាយ រួម 2. ទូទៅ ។ តែ ចំពោះ អ្នក សួរ ចម្លើយ គឺ នយោបាយ ខុស គ្នា នេះ ជា នយោបាយ ជំនាន់ នោះ ។
ROVER	1. រួម ទាំង កង សួរ ចម្លើយ រួម ទាំង ខាង ព្រៃ ស គឺ ជា វគ្គ នយោបាយ រួម ទូទៅ ។ 2. តែ ចំពោះ អ្នក សួរ ចម្លើយ គឺ នយោបាយ ខុស គ្នា នេះ ជា នយោបាយ ជំនាន់ នោះ ។

Table 3: Statistics of data reduction by the alignment based on the English sentences

Source	**Text sentence**	**Speech utterance**	**Target language speech**	
			utterance	**utterance**
EN	**82,078**	$79,857(-3\%)$	KM: $78,063(-5\%)$	FR: $78,016(-5\%)$
FR	$78,981(-4\%)$	$75,616(-4\%)$	KM: $73,967(-6\%)$	EN: $75,461(-4\%)$
KM	$80,417(-2\%)$	$65,679(-18\%)$	EN: $65,391(-19\%)$	FR: $64,203(-20\%)$

Second, the alignment between translated and ground truth was conducted using dynamic programming (DP) in a monotonic manner. Sentence boundary tokens were inserted following 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.

Fig. 1a shows that the alignment requires the MT to translate from source to target language. The alignment between English-French is acceptable because of the high translation quality of English-French. However, the translation quality of English/French-to-Khmer is limited; thus, the alignment still needs improvement. To address this problem, we applied the ROVER method, which will be described in Subsection 4.1.1, to combine the aligned Khmer text of English-Khmer and French-Khmer translations. With this voting result, we improved the performance by 0.5% of character error rate (CER) as shown in Table 1 and the example is given in Table 2. As a result, we obtained $82,078$ sentences in English aligned with $78,981$ sentences in French and $80,417$ sentences in Khmer, which means that only 4% and 2% in French and Khmer were discarded, respectively as presented in the second column of Table 3.

2.3.2. *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[d] using the Basic Expressions Travel Corpus[30] that was used in[31]. 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. The starting and ending timestamps of each sentence are aligned with a short audio data segment. At this stage, the alignment algorithm in Subsection 2.3.1 was used to generate the ground-truth text with the corresponding timestamp.

In this step, the performance of the ASR system is affected to the alignment result, which means that the better ASR performance will generate better alignment output. In this case, the text-to-speech alignment of English and French is well performed. As presented in the third column of Table 3, it reduced only 3% and 4% of English and French utterances, respectively. However, it reduced about 18% Khmer utterances by the text to speech alignment. The reason of this large reduction is the Khmer ASR model performance[31] was insufficient for transcribing some parts of the speech in this dataset. This is related to the domain and speaking-style mismatch, as the model was trained on traveling domain and reading style.

2.3.3. *Data cleaning*

For a usable corpus, we first cleaned the text data. We focused on the transcribed text that corresponds to speech data using the following process: removing unrelated parts that do not correspond to speech such as page headings, descriptions of the activity, and 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 changed to lowercase. For Khmer, we deleted the non-standard characters, punctuation marks, and other Latin symbols. We also normalized the text by correcting the spelling and following the order of the Khmer characters or diacritics, as presented in[32]. The numbers and abbreviations were also replaced by their standard spoken equivalent in all languages.

Second, we cleaned the speech corpus to ensure that the length of each audio segment was usable in ASR or ST. A usable length is in a range from $3s$ to $30s$. Each sentence of the transcription 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.

With the cleaning process, only a small portion $(1 - 2\%)$ of the original segmented speech utterances in the third column was reduced to the fourth and fifth columns of Table 3. There are two main reasons for this reduction of utterances:

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

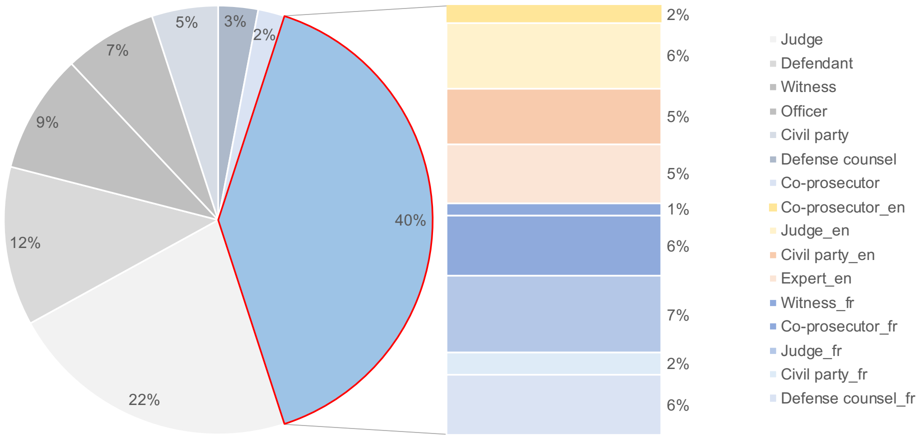


Fig. 2: Speaker distribution in a trilingual SLT corpus of the ECCC

i) mismatch between source speech and target text, and ii) long speech utterances which were not transcribed and segmented in the process of Section 2.3.2.

2.3.4. *Trilingual corpus statistics*

The graph in Fig. 2 shows the speaker distribution for each speaker group. Overall, 60%, 18%, 22% of speech is the original speech of Khmer, English, and French speakers, respectively. For Khmer source speech, 60% of speech is the original speech of Khmer speakers, including judges, defendants, witnesses, officers, co-prosecutors, defense counsels, and civil parties. The largest percentage is speech of the judges, which makes up 22% of the corpus, followed by 12% from the defendant, 9% from the witnesses, and 17% in total from other speakers. The remaining 40% of the speech is that of interpreters who interpreted the speech from native English and French speakers such as co-prosecutors, judges, civil parties, experts, witnesses, and defense counsels. For English source speech, 18% of speech is the original speech of English speakers, while the other 82% is the speech of interpreters who interpreted from the native of French and Khmer. Similarly, French has 22% of speech of native speakers, whereas another 78% is interpreted from English and Khmer native speakers.

Table 3 gives only the statistics of bilingual SLT, which cannot be used in some tasks such as multi-source translation or parallel joint training. Thus we selected the subset of the trilingual SLT corpus as shown in Table 4. This table gives the statistics of the SLT corpus of six-direction between Khmer, English, and French languages. It is approximately 146, 148, and 125 hours of speech in Khmer, English, and French, respectively, about 62K utterances of six directions. In terms of text, it is approximately 1.6M, 1.2M, and 1.3 words and the vocabulary sizes are 9K, 14K, and 20K in Khmer, English, and French, respectively. Finally, each language pair

Table 4: Statistics of each source language in the trilingual ECCC SLT corpus

Source	**#utterances**	**#words**	**vocabulary**	**#hour (train/dev/test)**
KM	62K	1.6M	9K	132/7/7
EN		1.2M	14K	134/7/7
FR		1.3M	20K	113/6/6

was split into training, development, and test sets, which are used in all experiments in this work.

3. Baseline end-to-end systems

The Transformer[4] is a recently state-of-the-art model applied in many fields, including applications of speech and language processing such as ASR, MT, and ST, also involved in this work. This architecture mainly stacks data input, encoder, decoder, and output building blocks. The data input building block uses embedding and position-encoding layers to transform an encoder's input sequences or features. On the other side, the output building block uses linear and softmax layers to generate the sequence of output tokens. Mainly, the encoder and decoder modules are core components that use the self-attention mechanism to calculate the attention score of each input sequence. Scaled dot product attention is then used to compute a weighted sum of values for a queries (**Q**) matrix of the three inputs: **Q**, keys (**K**) and values (**V**) as defined:

$$\text{Attention}(Q,K,V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

The encoder module is comprised of stacking the multi-head self-attention (MHA) and fully connected feed-forward network, coupled with layer normalization and residual connection. The attention module splits its **Q**, **K**, and **V** parameters N-ways and passes each split independently through a separate head. And all heads will be then combined to produce a final attention score using a concatenation operation:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1, .., \text{head}_\text{h})W^O, \tag{2}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \tag{3}$$

and the fully connected feed-forward network consists of two linear transformations with Rectified Linear Unit (ReLU) activation in between:

$$x = \text{FeedForward}(x), \tag{4}$$

$$\text{FeedForward}(x) = \max(0, xW_1 + b_1)W_2 + b_2, \tag{5}$$

The decoder has similar architecture as the encoder, which stacks multi-head attention with feed-forward networks in each layer. However, there are two multi-head attention sub-layers: i) a decoder self-attention, in which each position attends to all previous positions including the current position, and ii) encoder-decoder

Table 5: Word error rate (WER) of the ASR models in Khmer, English, and French; '*' model is used in cascade-ST and E2E-ST

Transformer ASR Model	WER		
	KM	EN	FR
w/o augmentation	23.6	6.9	14.5
w/ speed perturbation (SP)	22.2	6.6	14.0
w/ SpecAugment (SA)	21.8	6.4	13.8
w/ SP + SA *	**21.4**	**6.0**	**12.6**

attention, in which each position of the decoder attends to all positions in the last encoder layer.

Even though there has been a lot of interest recently in applying Transformer in speech and language processing to archive the promising results in both quality and efficiency, the task is limited to languages with large enough resources. This means that so far in low-resource language, the performance is limited due to data scarcity, resource quality, domain variability, and so on. Additionally, speaker-variability, speaking styles, and audio recorded environment[33] also affect the ASR performance. In contrast, the text style of written and spoken forms and text-speech mismatch and error propagation[34] generally influence to the MT and ST performance.

We conducted ASR, MT, and ST experiments using a Transformer-based[4] architecture implemented in ESPnet[35]. In all experiments, the network is comprised of six encoder layers and six decoder layers. The dimension of the feed-forward network was set to $2,048$, and the 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.

3.1. *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 with 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[36] as the vocabularies for each language. Speech perturbation[37] and SpecAugment[38] were applied as data augmentation. All system performances are evaluated in WER and shown in Table 5. The table shows that English ASR performs better compared to other languages. Its WER is 6.0% followed by French with 12.6% and the performance of the Khmer language is worst with 21.4%. The Khmer speech is the most challenging in this corpus because the

Table 6: BLEU for translation of each language pair in a TriECCC corpus

Source	**Target**	**BLEU**		
		MT	**Cascade-ST**	**E2E-ST**
KM →	EN	16.63	15.14	13.81
	FR	11.53	10.66	9.39
EN →	KM	14.44	14.15	14.14
	FR	25.01	24.32	20.83
FR →	KM	10.54	9.82	10.26
	EN	27.37	25.17	23.64

original Khmer speech was spoken by the older people who were the victims of the Khmer Rouge regime. Most of them are illiterate in the Khmer language. They sometimes cannot pronounce words correctly, and exhibit disfluency and emotions in their speech during the trial as mentioned in[20]. On the other hand, 78% speech of English and 82% speech of French were spoken by middle-age interpreters and other well-prepared speakers, including judges, co-prosecutors, civil parties, and so on.

3.2. *Machine translation (MT)*

For MT, we trained another Transformer-based model for 100 epochs with a batch size of 96. However, the model tends to converge within 50 epochs. The Noam optimizer was used with 8K warmup steps and an initial learning rate of 1. All punctuation marks were stripped and converted to lowercase English and French in each language pair. We applied 15K BPE tokens of trilingual vocabularies, which resulted in 5K per language. The translation performances are reported using BLEU [39], as shown in Table 6.

The translation between English and French performs much better than that between Khmer and English/French. This is because of the disfluency of Khmer transcription, which was transcribed from the disfluent speech of the original Khmer speakers. Moreover, the translations between Khmer and English perform better than in Khmer and French. This is reasonable because English was directly used as the source language for the bilingual sentence alignment to Khmer and French, which were indirectly aligned.

3.3. *Speech translation (ST)*

The E2E-ST front-end configuration is similar to the ASR system. The speed perturbation and SpecAugment were applied as the speech data augmentation. The 15K BPE tokens of trilingual vocabulary were used as they were for MT. Note that the trilingual vocabulary was used for all translation models because it is useful for transfer learning purpose on both ST and MT in this work. In ST systems, we trained only 60 epochs with a batch size of 64. The ASR and MT pre-trained models, which were presented in the previous Sections 3.1 and 3.2, were respectively used to initialize the E2E-ST encoder and decoder. With this initialization, the E2E-ST can achieve reasonable performance, as described in[40]. For cascade-ST, we first transcribed the speech using the ASR system, and then this output text was fed to the MT model to translate into the target language. The results are reported in Table 6.

The table shows the performance of both E2E-ST and cascade-ST. Overall, the cascade-ST system has a slightly lower BLEU score compared to MT system, but it is better than E2E-ST in most cases. Generally, the ST performance has a big problem in non-monotonic alignment of speech-text or text-speech, that is why their performances were worse than the normal MT models. Moreover, the speech condition is also an influential factor on ST performance, for instance, the translation to Khmer by the E2E-ST system is comparable or better than cascade-ST models. This is because the English and French ASR performance is better than the Khmer ASR performance.

4. Enhancement of MT

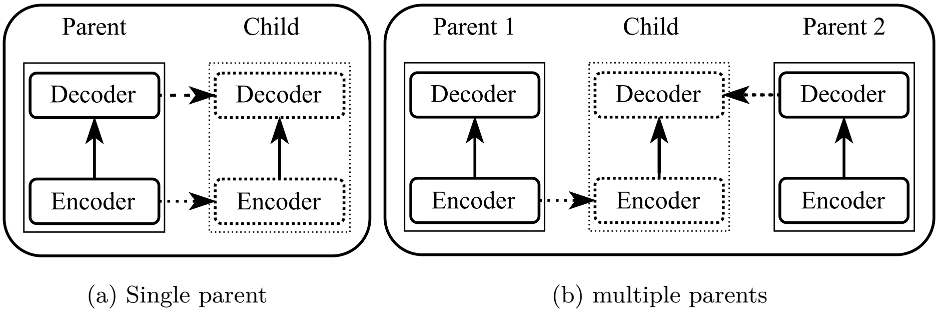
4.1. *Methods*

In order to enhance ASR and MT of low-resource languages, many approaches have been investigated including multilingual training, system combination, transfer learning, and knowledge distillation. In this work, we focus on improving the MT of the Khmer language from/to English and French by using a system combination of ROVER and cross-lingual transfer learning methods.

4.1.1. *ROVER method*

ROVER is one of the most commonly used methods to combine the hypotheses of multiple ASR outputs in system combination. Originally, ROVER performs two-step procedures composed of word alignment and voting mechanisms. Word alignment combines the multiple outputs using dynamic programming to a minimal-cost word transition network (WTN). Then, the voting mechanism selects the best output word sequence based on the frequency of occurrence and word-level confidence score. This method has been shown to significantly reduce the WER[19]. However, the voting result will be poor if the confidence score of each output system is not reliable. Moreover, the voting result will not outperform the individual system if multiple systems do not have complementary errors[41].

In this work, we combine only two translation systems which produce different hypotheses of the same target language from different language source input of the same content. Specifically, we used the ROVER method to combine the translation output of English-to-Khmer and French-to-Khmer MT systems to enhance the hypothesis of the Khmer language.



(a) Single parent (b) multiple parents

Fig. 3: The fine-tuning process: (a) the single parent, (b) the multiple parents

4.1.2. *Transfer learning*

The transfer learning methods have been successfully applied to applications in speech and language processing, including speech recognition, document classification, and sentiment analysis[42], MT[43] and various downstream tasks[44,45]. In MT task, using the pre-trained model of high-resource language pair (e.g., Spanish-to-English) is effective in assisting a low-resource language pair (e.g., Catalan-to-English)[43]. With this approach, a parent model is trained on a high-resource language pair, and then the trained parameters are used to initialize a child model, which is trained on the desired low-resource language pair. On the other hand, a multiple parents fine-tuning process[46], which has two parents to transfer to a child model in two steps, is beneficial when multiple languages are involved. For instance, to improve a child model (e.g., German-to-Czech), we can first use a parent (e.g., German-to-English) to initialize to encoder parameters, and another parent (e.g., English-to-Czech) to initialize the parameters of the decoder of the child model. We can transfer some or all parameters from the parent to a child model at the initializing stage. However, the effectiveness of fine-tuning might be different when transfer learning is conducted in different parameters or layers, especially with a complex model with multiple modules such as Transformer.

In this work, we use English-to-French and French-to-English MT systems as the pre-trained models because we aim to leverage the well-trained model of the high-resource language pairs. Moreover, as presented in Table 6, the English-French models show much better translation quality than Khmer from/to English and French.

We investigate initialization from both single and multiple parents as shown in Fig 3. We first investigate the use of a single parent (Fig 3a) to initialize encoder, decoder, and both encoder and decoder modules (e.g., the English-to-French model is used to initialize the English-to-Khmer and Khmer-to-French models). Secondly, we conduct the initialization from multiple parents as in Fig 3b (e.g., the encoder part is fine-tuned from the English-to-French model, and the decoder part is fine-tuned from the French-to-English model or vice versa).

Table 7: Result of ROVER method for MT outputs of Khmer

Source	**Target**	**Baseline**	**ROVER**
EN	KM	14.44	**14.79**
FR		10.54	

For the fine-tuning from the pre-trained model, we used the same configuration of the original MT as presented in Subsection 3.2. However, we trained each model only 30 epochs, and the models were well converged. In each fine-tuning process, the initializing was applied to all layers or some of the specific layers in the encoder-decoder modules of the Transformer, but initializing to all layers of both encoder and decoder shows the best performance.

4.2. *Experimental evaluations*

4.2.1. *Voting the Khmer translation using ROVER method*

Table 7 shows that the ROVER method improved the translation to Khmer. Specifically, it outperforms the BLEU score of English-to-Khmer and French-to-Khmer systems by 0.35 and 4.25, respectively. This is because the ROVER method increases the variety of output by combining the two hypotheses.

4.2.2. *Fine-tuning the Khmer translation using pre-trained model*

Table 8 presents the best practice of initializing with single and multiple parents by transferring the parameters to the encoder, decoder only, or both encoder and decoder modules. In the single-parent case, the English-to-French model is used to initialize the English-to-Khmer and Khmer-to-French models, while the French-to-English model is used to initialize the French-to-Khmer and Khmer-to-English models. In the multiple parents case, the English-to-French model is used for the encoder module and French-to-English model is used for the decoder side or vice versa.

Table 9 compares the performance of the fine-tuning approach, which uses the pre-trained model to initialize all layers of encoder, decoder, or encoder-decoder modules, compared with the baseline performance. Generally, initializing with the pre-trained model is effective for boosting the MT performance in both directions of Khmer MT systems. Transferring the knowledge to the encoder only is usually better than the decoder, but initializing both encoder and decoder shows the best improvement in all systems. Additionally, the single-parent fine-tuning shows better performances in most models, except for the English-to-Khmer.

In terms of Khmer as a source language, using the pre-trained model of the same target language gives the best performance. Specifically, the pre-trained model of French-to-English improved the Khmer-to-English MT, whereas the English-to-French model improved the Khmer-to-French MT. This is because the pre-trained

Table 8: The best practice in the use of the pre-trained model to initialize each Khmer MT model (Enc.-Dec.: Encoder-Decoder)

Source	**Target**	**Pre-trained parent models used for initialization**			
		Encoder	**Decoder**	**Enc.-Dec.**	**Multiple parents**
KM	EN	**FR**-EN	FR-**EN**	**FR**-**EN**	**EN**-FR and FR-**EN**
	FR	**EN**-FR	EN-**FR**	**EN**-**FR**	**FR**-EN and EN-**FR**
EN	KM	**EN**-FR	EN-**FR**	**EN**-**FR**	**EN**-FR and FR-**EN**
FR		**FR**-EN	FR-**EN**	**FR**-**EN**	**FR**-EN and EN-**FR**

Table 9: Comparison the best performance of fine-tuning approach initializing the pre-trained model into encoder, decoder, or both on each Khmer MT model

Source	**Target**	**Performance of each initial option (BLEU↑)**				
		Baseline	**Encoder**	**Decoder**	**Enc.-Dec.**	**Multiple**
KM	EN	16.63	17.56	17.04	**18.16**	17.64
	FR	11.53	12.69	12.45	**13.77**	13.61
EN	KM	14.44	15.37	15.11	15.54	**15.61**
FR		10.54	11.81	11.32	**12.13**	11.85

model helps to generalize the alignment from Khmer to the target languages. Especially, the decoder part can be enhanced from the translation knowledge of the pre-trained model in the same target language.

On the other hand, when Khmer is a target language, the French-to-English model was the best pre-trained model to enhance the performance of the MT performance. In this case, the use of this pre-trained model to initialize both encoder and decoder increased the performance with the single-parent fine-tuning process. Whereas using the English-to-French pre-trained model to initialize the encoder and initializing decoder with the French-to-English model shows the best performance in English-to-Khmer MT model. There are two main reasons for this improvement, i) using the same source language between the new and pre-trained models is helpful in the alignment process, ii) the pre-trained model of French-to-English has the better performance than English-to-French.

Overall, using the same source or target between the new and pre-trained MT models can enhance the performance of low-resource MT systems because the knowledge of the pre-trained model improved the alignment between the source and target languages. As a result, the fine-tuning process improved the BLEU score by 2.24 and 1.59 points for Khmer-to-French and French-to-Khmer, respectively. On the other hand, the translations between Khmer and English were improved by only 1.53 and 1.17 points for Khmer-to-English and English-to-Khmer, respectively.

Table 10 and 11 show examples of the compared methods in the translated result between Khmer and other languages. The output of the transfer learning shows consistent improvement as it gives a complete sentence with the same meaning, whereas the ROVER method sometimes generated an incomplete sentence, or

Table 10: Examples of the comparison of all methods in English-Khmer MT models, the *italic text* in the "()" is the translated text in to English.

Khmer to English	
Hypothesis	ខ្ញុំ មិន ដឹង ជា អក្សរ អ្នក ណា ទេ សរសេរ មក ដើម្បី ឲ្យ ខ្ញុំ សួរ បន្ថែម ទៀត
Reference	i am not sure who annotated this confession in order for me to further interrogate the person
Baseline	i don't know who was the commander of the battalion so that i can ask further questions
Encoder init.	i don't know who was the chief of the battalion so that i can interrogate further
Decoder init.	i did not know who the circular or who was from the battalion to provide further interrogation
Enc.-Dec. init.	i was not sure who was the chief of the unit in order to ask for further questions
English to Khmer	
Hypothesis	i am not sure who annotated this confession in order for me to further interrogate the person
Reference	ខ្ញុំ មិន ដឹង ជា អក្សរ អ្នក ណា ទេ សរសេរ មក ដើម្បី ឲ្យ ខ្ញុំ សួរ បន្ថែម ទៀត *(I don't know who wrote the letter in order for me to ask more questions)*
ROVER	ខ្ញុំ មិន ប្រាកដ ថា អ្នក ណា ជា អ្នក ចារ លើ សេចក្ដី សារភាព នេះ ដើម្បី ឲ្យ ខ្ញុំ អត់ សួរ បន្ថែម ទៀត *(I don't know who wrote this confession in order me to not ask more questions)*
Baseline	ខ្ញុំ មិន ប្រាកដ ថា អ្នក ណា ជា អ្នក ចារ លើ សេចក្ដី សារភាព នេះ ដើម្បី ឲ្យ ខ្ញុំ សួរ បន្ថែម ទៀត *(I am not sure who wrote this confession in order me to ask more questions)*
Encoder init.	ខ្ញុំ មិន ដឹង ថា អ្នក ណា ជា អ្នក ចារ លើ សេចក្ដី សារភាព នេះ ដើម្បី ឲ្យ ខ្ញុំ សួរ បន្ត ទៀត ទេ *(I am not sure who wrote this confession in order me to continue to ask more questions)*
Decoder init.	មិន ដឹង ជា អក្សរ អ្នក ណា ទេ សរសេរ មក ដើម្បី ឲ្យ ខ្ញុំ សួរ បន្ថែម ទៀត *(Don't know who wrote in order me to ask more questions)*
Enc.-Dec. init.	ខ្ញុំ មិន ប្រាកដ ថា អ្នក ណា សរសេរ ចម្លើយ សារភាព នេះ ដើម្បី ឲ្យ ខ្ញុំ សួរ បន្ថែម ទៅ លើ អ្នក ទោស នោះ ទេ *(I am not sure who wrote this confession in order me to ask the prisoners more questions)*

changed the meaning of the output sentence because this method copies different words from another system outputs.

5. Conclusions

In this work, we created the largest-ever simultaneous SLT corpus from the ECCC dataset of 222 sessions for six directions in Khmer, English, and French. We kept a large proportion of the original dataset by using monotonic sentence alignment and word-based distance calculation. This alignment requires the segmentation of the sentences in the source language only. This method is very effective and helpful

Table 11: Examples of the comparison of all methods in French-Khmer MT models, the *italic text* in the "()" is the translated text in to English.

Khmer to French	
Hypothesis	ខ្ញុំ មិន ដឹង ជា អក្សរ អ្នក ណា ទេ សរសេរ មក ដើម្បី ឲ្យ ខ្ញុំ សួរ បន្ថែម ទៀត
Reference	je ne sais pas qui a écrit ceci en tout cas cela me donnait l'ordre
Baseline	je ne sais pas s'il y avait quelqu'un avec quelqu'un pour qu'il
Encoder init.	je ne sais pas qui était le chef du bataillon pour lui permettre de poser des questions
Decoder init.	je ne sais pas qui était le chef de la prison pour lui permettre de poser des questions supplémentaires
Enc.-Dec. init.	je ne sais pas qui était le chef du bataillon pour qu'il poursuive ses questions
French to Khmer	
Hypothesis	je ne sais pas qui a écrit ceci en tout cas cela me donnait l'ordre
Reference	ខ្ញុំ មិន ដឹង ជា អក្សរ អ្នក ណា ទេ សរសេរ មក ដើម្បី ឲ្យ ខ្ញុំ សួរ បន្ថែម ទៀត *(I don't know who wrote the letter in order for me to ask more questions)*
Baseline	ខ្ញុំ មិន ដឹង ថា អ្នក ណា បញ្ជា ខ្ញុំ អត់ បាន ដឹង ទេ *(I don't know who ordered me, I don't know)*
ROVER	ខ្ញុំ មិន ប្រាកដ ថា អ្នក ណា ជា អ្នក ចារ លើ សេចក្តី សារភាព នេះ ដើម្បី ឲ្យ ខ្ញុំ អត់ សួរ បន្ថែម ទៀត *(I don't know who wrote this confession in order me to not ask more questions)*
Encoder init.	ខ្ញុំ មិន ដឹង ថា អ្នក ណា ជា អ្នក បញ្ជា ខ្ញុំ អត់ ដឹង ទេ *(I don't know who ordered me, I don't know)*
Decoder init.	ខ្ញុំ មិន ដឹង ថា អ្នក ណា ជា អ្នក បញ្ជា ឲ្យ ខ្ញុំ ធ្វើ *(I don't know who ordered me to do that)*
Enc.-Dec. init.	ខ្ញុំ មិន ដឹង ថា អ្នក ណា ជា អ្នក បញ្ជា ឲ្យ ខ្ញុំ ធ្វើ យ៉ាង ណា នោះ ទេ *(I don't know who ordered me to do something)*

in aligning a rich-resource language to other low-resource languages. Finally, we built the 146, 148, and 125 hours in length of speech and 1.6, 1.2, and 1.3 million words in the text of Khmer, English, and French, respectively. Furthermore, we conducted E2E ASR, MT, and ST experiments on the constructed corpus and obtained reasonable performance.

To improve the Khmer MT, we conducted ROVER and fine-tuning the pre-trained models of English-to-French and French-to-English. The results show that the ROVER is practical for combining the systems with similar performance. Meanwhile, the use of the pre-trained model was effective in improving the BLEU score. Initializing both encoder and decoder modules is most effective.

This corpus will be useful for speech and language research of the Khmer language. It will be helpful for many kinds of applications in speech and language processing research, including ASR, MT, and ST, and multi-lingual or multi-source ASR, MT, and ST or even speaker recognition as presented in[23]. Moreover, this alignment method will benefit similar datasets such as meetings, classroom lectures, and TV programs.

References

1. A. Graves, S. Fernandez, F. Gomez and J. Shmidhuber, Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks, in *Proceedings of ICML*, 2006.
2. A. Graves, Sequence Transduction with Recurrent Neural Networks, *Proceedings of ICML* (2012).
3. W. Chan, N. Jaitly, Q. Le and O. Vinyals, Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, in *Proceedings of ICASSP*, (IEEE, 2016).
4. A. Vaswani, N. S. abd Niki Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, Attention is All You Need, in *Proceedings of NeurIPS*, 2017.
5. A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu and R. Pang, Conformer: Convolution-augmented Transformer for Speech Recognition, in *Proceedings of Interspeech*, 2020.
6. M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri and M. Turchi, MuST-C: a Multilingual Speech Translation Corpus, in *Proceedings of NAACL-HLT*, 2019.
7. M. Post, G. Kumar, A. Lopez, D. Karakos, C. Callison-Burch and S. Khudanpur, Improved speech-to-text translation with the Fisher and Callhome Spanish–English speech translation corpus, in *Proceedings of IWSLT*, 2013.
8. 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 *Proceedings of ICASSP*, 2020.
9. E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D. Oard and M. Post, The multilingual tedx corpus for speech recognition and translation, *ArXiv* **abs/2102.01757** (2021).
10. F. J. Och and H. Ney, Statistical multi-source translation, in *Proceedings of MT Summit*, Citeseer2001.
11. E. Garmash and C. Monz, Ensemble learning for multi-source neural machine translation, in *Proceedings of COLING*, 2016, pp. 1409–1418.
12. B. Zoph and K. Knight, Multi-source neural translation, in *Proceedings of NAACL-HLT*, June 2016, pp. 30–34.
13. R. Dabre, F. Cromieres and S. Kurohashi, Enabling multi-source neural machine translation by concatenating source sentences in multiple languages, *Proceedings of MT Summit* (2017).
14. Y. Nishimura, K. Sudoh, G. Neubig and S. Nakamura, Multi-source neural machine translation with data augmentation, *arXiv preprint arXiv:1810.06826* (2018).
15. Y. Nishimura, K. Sudoh, G. Neubig and S. Nakamura, Multi-source neural machine translation with missing data, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2020) 569–580, doi:10.1109/TASLP.2019.2959224.
16. Z. Lu, X. Li, Y. Liu, C. Zhou, J. Cui, B. Wang, M. Zhang and J. Su, Exploring multi-stage information interactions for multi-source neural machine translation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2021) doi: 10.1109/TASLP.2021.3120592.
17. X. Zhou, E. Yılmaz, Y. Long, Y. Li and H. Li, Multi-Encoder-Decoder Transformer for Code-Switching Speech Recognition, in *Proceedings of Interspeech*, 2020.
18. Y.-F. Cheng, H.-S. Lee and H.-M. Wang, AlloST: Low-Resource Speech Translation Without Source Transcription, in *Proceedings of Interspeech*, 2021, pp. 2252–2256.
19. J. Fiscus, A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover), in *Proceedings of ASRU*, 1997.
20. K. Soky, M. Mimura, T. Kawahara, S. Li, C. Ding, C. Chu and S. Sam, Khmer

Speech Translation Corpus of the Extraordinary Chambers in the Courts of Cambodia (ECCC), in *Proceedings of O-COCOSDA*, 2021.

21. F. E. Huffman, Cambodian System of Writing and Beginning Reader, *Yale University Press* (1970).
22. T. Nakazawa, N. Doi, S. Higashiyama, C. Ding, R. Dabre, H. Mino, I. Goto, W. P. Pa, A. Kunchukuttan, Y. Oda, S. Parida, O. Bojar and S. Kurohashi, Overview of the 6th workshop on Asian translation, in *Proceedings of ACL*, November 2019.
23. K. Soky, S. Li, M. Mimura, C. Chu and T. Kawahara, On the use of speaker information for automatic speech recognition in speaker-imbalanced corpora, in *Proceedings of APSIPA ASC*, 2021.
24. F. Braune and A. Fraser, Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora, in *Proceedings of COLING*, 2010.
25. C. Dyer, V. Chahuneau and N. A. Smith, A simple, fast, and effective reparameterization of IBM model 2, in *Proceedings of NAACL-HLI*, 2013.
26. B. Thompson and P. Koehn, Vecalign: Improved sentence alignment in linear time and space, in *Proceedings of EMNLP-IJCNLP*, 2019.
27. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst, Moses: Open source toolkit for statistical machine translation, in *Proceedings of ACL*, 2007.
28. T. Kiss and J. Strunk, Unsupervised multilingual sentence boundary detection, *Comput. Linguist.* (December 2006) p. 485–525, doi:10.1162/coli.2006.32.4.485.
29. V. Chea, Y. K. Thu, C. Ding, M. Utiyama, A. Finch and E. Sumita, Khmer word segmentation using conditional random fields, in *Proceedings of Khmer Natural Language Processing (KNLP)*, 2015.
30. G. Kikui, E. Sumita, T. Takezawa and S. Yamamoto, Creating corpora for speech-to-speech translation, in *Proceedings of EUROSPEECH*, 2003.
31. K. Soky, S. Li, T. Kawahara and S. Seng, Multi-lingual transformer training for khmer automatic speech recognition, in *Proceedings of APSIPA ASC*, 2019.
32. B. Marie, H. Kaing, A. M. Mon, C. Ding, A. Fujita, M. Utiyama and E. Sumita, Supervised and unsupervised machine translation for Myanmar-English and Khmer-English, in *Proceedings of the 6th Workshop on Asian Translation*, 2019.
33. U. Shrawankar and V. Thakare, Adverse conditions and asr techniques for robust speech user interface (2013).
34. J. Niehues, E. Salesky, M. Turchi and M. Negri, Tutorial proposal: End-to-end speech translation, in *Proceedings of ACL*, April 2021, pp. 10–13.
35. 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 *Proceedings of Interspeech*, 2018.
36. R. Sennrich, B. Haddow and A. Birch, Neural machine translation of rare words with subword units, in *Proceedings of ACL*, 2016.
37. T. Ko, V. Peddinti, D. Povey and S. Khudanpur, Audio Augmentation for Speech Recognition, in *Proceedings of Interspeech*, 2015.
38. 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 *Proceedings of Interspeech*, 2019.
39. K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in *Proceedings of ACL*, July 2002, pp. 311–318.
40. H. Inaguma, S. Kiyono, K. Duh, S. Karita, N. Yalta, T. Hayashi and S. Watanabe, ESPnet-ST: All-in-one speech translation toolkit, in *Proceedings of ACL*, 2020.
41. O. Siohan, B. Ramabhadran and B. Kingsbury, Constructing ensembles of asr systems

using randomized decision trees, in *Proceedings. ICASSP*, 2005.

42. D. Wang and T. F. Zheng, Transfer learning for speech and language processing, *CoRR* **abs/1511.06066** (2015) http://arxiv.org/abs/1511.060661511.06066.
43. B. Zoph, D. Yuret, J. May and K. Knight, Transfer learning for low-resource neural machine translation, in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, (Association for Computational Linguistics, Austin, Texas, November 2016), pp. 1568–1575.
44. J. Devlin, M. Chang, K. Lee and K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in *Proceedings of NAACL-HLT*, eds. J. Burstein, C. Doran and T. Solorio2019, pp. 4171–4186.
45. 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 NAACL-HLT*, 2019, pp. 58–68.
46. Y. Kim, P. Petrov, P. Petrushkov, S. Khadivi and H. Ney, Pivot-based transfer learning for neural machine translation between non-English languages, in *Proceedings of EMNLP-IJCNLP*, November 2019, pp. 866–876.