Interspeech 2022
18-22 September 2022, Incheon, Korea

Leveraging Simultaneous Translation for Enhancing Transcription of Low-resource Language via Cross Attention Mechanism

Kak Soky[1,2], *Sheng Li*[2], *Masato Mimura*[1], *Chenhui Chu*[1], *Tatsuya Kawahara*[1]

[1]Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto, Japan
[2]National Institute of Information and Communications Technology (NICT), Kyoto, Japan
{soky,mimura,kawahara}@sap.ist.i.kyoto-u.ac.jp

Abstract

This work addresses automatic speech recognition (ASR) of a low-resource language using a translation corpus, which includes the simultaneous translation of the low-resource language. In multi-lingual events such as international meetings and court proceedings, simultaneous interpretation by a human is often available for speeches of low-resource languages. In this setting, we can assume that the content of its back-translation is the same as the transcription of the original speech. Thus, the former is expected to enhance the later process. We formulate this framework as a joint process of ASR and machine translation (MT) and implement it with a combination of cross attention mechanisms of the ASR encoder and the MT encoder. We evaluate the proposed method using the spoken language translation corpus of the Extraordinary Chambers in the Courts of Cambodia (ECCC), achieving a significant improvement in the ASR word error rate (WER) of Khmer by 8.9% relative. The effectiveness is also confirmed in the Fisher-CallHome-Spanish corpus with the reduction of WER in Spanish by 1.7% relative.[1]

Index Terms: automatic speech recognition, machine translation, multi-lingual corpus, low-resource language, Khmer

1. Introduction

While deep learning, particularly end-to-end (E2E) modeling [1, 2, 3, 4, 5] has significantly advanced automatic speech recognition (ASR), ASR of low-resource languages still remains one of the big challenges as these languages do not have a sufficient amount of training data. Another task for low-resource languages is machine translation (MT) or spoken language translation (SLT) because many foreign people do not understand these languages. In international meetings such as UN conventions [6] and EU Parliaments [7], simultaneous interpretation by human translators is often available [8]. In this work, we use an international court proceedings of the Extraordinary Chambers in the Courts of Cambodia (ECCC) [9], in which Khmer is the primary language and English and French translations are available. MT and SLT corpora have been built on these types of datasets.

In this study, we focus on ASR of low-resource languages (e.g. Khmer), which is also the basis of SLT of these languages, by leveraging the translation corpus. Here we assume ASR of fluent speech of the human translators in a resource-rich language (e.g. English and French) is perfect, thus use the output text instead of speech. Note that transcription of the original speech (i.e. Khmer) is still mainly required for the Khmer people. In this setting, the content of back-translation of the translation text (e.g. English-to-Khmer) must be the same as the transcription of the original speech (i.e. Khmer). Therefore, the former is expected to enhance the later, specifically, MT output is expected to complement the ASR process. This is analogous to a scenario in which we (e.g. Japanese) can more easily recognize a foreign-language (e.g. English) movie with simultaneous subtitles of the native language (e.g. Japanese).

In previous studies, multi-task learning and system combination of multiple models have been investigated to improve ASR performance, for example, the integration between ASR and MT models trained in multiple iterative stages [10, 11]. This integration is also applied to computer-assisted translation application [12, 13, 14, 15, 16]. However, these works used independent systems of ASR and MT, similar to the idea in ROVER [17], which ensembles the output of multiple ASR recognizers using an alignment and then voting mechanism. Another approach is to train a large text-only or text-to-text model to be coupled with the ASR model. Wang et al. [18] trained a large decoder of text corpus to alleviate the need for an external language model. Yusuf et al. [19] trained a bank of shallow task-specific modality encoders including MT and mask language model (MLM) as the auxiliary task to ASR. These works require a large text corpus, which is not the case in low-resource languages.

In contrast, we propose a joint ASR-MT framework to enhance the ASR performance of a low-resource language using MT output. It trains ASR and MT modules using input sources of speech and its parallel translation text simultaneously. Our proposed method jointly trains dual encoders of ASR and MT together and then uses the translation knowledge from a rich-resource language to assist the transcription of a low-resource language via a cross-attention mechanism in a single E2E model [4]. Although the proposed method trains multiple encoders simultaneously, it is different from multi-source MT [20, 21, 22], which uses multiple inputs of text in different languages, and it is different from cascade speech translation (ST), which is stacking the ASR and MT systems, and the E2E-ST, which uses the ASR encoder and MT decoder. Our target is to improve ASR of the low-resource languages.

We first evaluate our proposed method using the multi-lingual SLT corpus of ECCC, in which the goal is to improve the transcription performance of the Khmer speech using the translation from English or French. We then apply this method to the Fisher-CallHome corpus [23] for improving the transcription of Spanish with the use of translation from English.

The rest of this paper is structured as follows. Section 2 gives a brief overview of the related work. We then present the detailed concept of our proposed method for jointly trained ASR and MT in Section 3. In Section 4, we describe the setup of the experiments, and present the experimental results. We finally conclude the paper with final remarks in Section 5.

[1] Source code is available at: https://github.com/ksoky/jointlytrained

10.21437/Interspeech.2022-343

2. Related work

The study of enhancing the ASR system on the target language of the human translator using the translation of the source document was investigated by Paulik et al. [10], who analyzed the effects of different MT models to be integrated into the ASR system in multiple iterations. In each iteration, they updated an n-gram language model for rescoring the ASR n-best list, whereas in [11], the ASR system was improved by extracting the MT n-best list in several iterations to rescore the ASR n-best list, where both ASR and MT were conducted in parallel. Similarly, Khadivi et al. [12, 13] also integrated MT and ASR models for computer-assisted translation. In these works, they used independent ASR and MT models and then interactively updated the n-gram language model of each system in multiple iterations or integrated the outputs of these systems.

Recently, Macháček et al. [8] compared quality and latency of spoken translation systems from English to Czech using Europarl Simultaneous Interpreting Corpus. This investigation showed that the interpreters tend to compress and simplify the speech, which means the translations keep the content but are not necessarily literal. Yusuf et al. [19] proposed a framework to improve ASR with a unified speech and text encoder-decoder, in which the system jointly trained an attention-based of ASR and a variety of text-to-text transduction tasks including MT and MLM. All tasks shared parameters of encoder layers and decoder modules, but the MT and MLM were trained on a large text corpus which is unpaired to the ASR corpus.

In this study, we enhance ASR of a low-resource language by jointly training the ASR and MT in a single E2E model using the paired data between audio-to-text for ASR and text-to-text for MT.

3. Proposed method

The tasks of ASR and MT are to generate a text from a source speech and from another language text, respectively. Therefore, we propose to jointly train these ASR and MT models in a single E2E model. Specifically, we incorporate the translation knowledge from a rich-resource language to enhance transcription of speech of a low-resource language.

Similar to multi-task learning, we conduct a joint training of both ASR and MT encoders as shown in Figure 1, in which an original speech in a language (L1, e.g. Khmer) and its corresponding translation in another language (L2, e.g. English) are used as the input sources. Note that we assume ASR of the translators' speech (fluent English/French) is perfect, thus use the transcription text instead of speech in this work[2]. We then combine the cross-attention of ASR and MT encoders to the joint decoder to improve automatic transcription. With this combination, the translation knowledge is used to enhance the transcription process.

This proposed framework formulates that, with a given set of speech utterances in L1, $\{X_1, X_2, ..., X_e\}$, and their translations in L2, $\{Z_1, Z_2, ..., Z_e\}$, the model predicts text transcription in L1, $\{Y_1, Y_2, ..., Y_e\}$, where e is the total number of sentences or utterances.

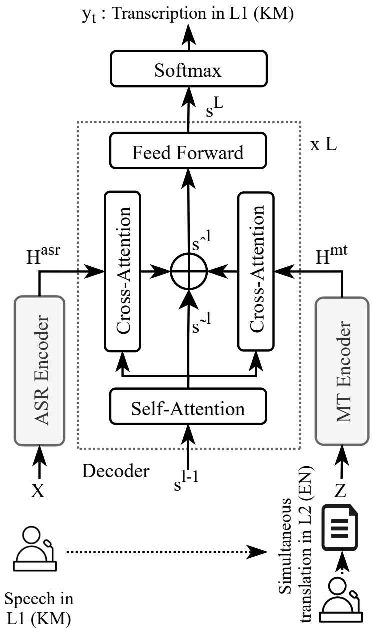

Figure 1: *Proposed method of joint ASR and MT*

3.1. Dual encoders

The proposed architecture comprises of both ASR and MT encoders. Each encoder is based on the Transformer architecture [4], but we train both encoders jointly in a single model. For each sequence of n acoustic features in L1, $X = \{x_1, x_2, ..., x_n\}$, and sequence of m tokens in L2, $Z = \{z_1, z_2, ..., z_m\}$, the encoders predict the intermediate representation matrices H^{asr} and H^{mt}.

$$\begin{aligned} H^{\text{asr}} &= \text{Encoder}(X), \\ H^{\text{mt}} &= \text{Encoder}(Z). \end{aligned} \tag{1}$$

3.2. Joint decoder

The decoder network is implemented as a stack of L modified Transformer layers. Unlike a standard Transformer decoder, each layer in our decoder has two distinct cross attention components in order to combine information from both of the ASR and MT encoders. More specifically, the output of each layer at the t-th decoding step $S_t^l = \{s_1^l, s_2^l, ..., s_t^l\}$ is calculated using the representation from the ASR encoder H^{asr} and that from the MT encoder H^{mt}, as well as the output of the previous decoder layer S_t^{l-1}. Note that we define s_j^0 as the embedding of the j-th predicted token y_i .
Each s_t^l is calculated as:

$$\tilde{s}_t^l = \text{Attention}(s_t^{l-1}, S_t^{l-1}, S_t^{l-1}), \tag{2}$$

$$\begin{aligned} \hat{s}_t^l = \text{Attention}(\tilde{s}_t^l, H^{\text{asr}}, H^{\text{asr}}) + \\ \text{Attention}(\tilde{s}_t^l, H^{\text{mt}}, H^{\text{mt}}) + \tilde{s}_t^l, \end{aligned} \tag{3}$$

$$s_t^l = \text{FeedForward}(\hat{s}_t^l). \tag{4}$$

Here, each self-attention component takes a query Q, key K and value V as the inputs, and its output is obtained as:

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{5}$$

[2] This assumption is not so unrealistic, given the WER of Librispeech is less than 3% [24]. Moreover, the number of translators is only three in this dataset, thus we can have similar performance when we train the speaker-adapted model.

Then, the output probability of the current token y_t is given as:

$$P(y_t|S_t^0, X, Z) = \mathrm{Softmax}(s_t^L) \tag{6}$$

Finally, the probability of the transcription text Y is defined as:

$$P(Y|X, Z) = \prod_{t=1} P(y_t|S_t^0, X, Z) \tag{7}$$

Note that without cross-attention from MT, the network is virtually the same as the standard Transformer-based ASR system. Thus, the proposed system is regarded as its extension.

3.3. Objective function

To optimize the model training, each task has a well-defined loss function. With the proposed method, there are two losses of ASR and MT, which could be optimized with multi-task learning. However, the output of these two tasks are essentially the same, and each loss is propagated to the respective model.

4. Experimental evaluations

4.1. Dataset

ECCC is a court dataset consisting of text and speech in Khmer, English, and French. This dataset has been built as a bilingual Khmer-English ECCC corpus for MT, which has only text data [25], a Khmer speech-to-text corpus for ASR [26], and an SLT corpus of the Khmer to English and French, reported in [9]. Our main target is to improve speech transcription by incorporating MT (English to Khmer or French to Khmer). The SLT corpus of 155 hours in length of speech and 1.7M words in text are used to conduct the experiments. For data preparation, we used the same process as that described in [9].

4.2. Model training

We implemented the model using a Transformer-based architecture of the ESPnet [27]. Following the standard setup, we used 80-dimensional log-melscale filterbank coefficients and 3-dimensional pitch features. Speech perturbation [28] and SpecAugment [29] were applied for speech data augmentation. The network is composed 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 with the dimension of 256. This network was started with down-sampling using a two-layer time-axis convolutional layer with 256 channels, stride size of 2, and kernel size of 3. The model was jointly trained with CTC (weight α = 0.3) for 45 epochs with a single 12-GB Titan X GPU using a batch size of 64. The "Noam" optimizer was used with $25,000$ warmup steps and an initial learning rate of 5. The byte pair encoding (BPE) [30] of the source and target languages was set to $5,000$ for each. We used a joint source and target vocabularies for the proposed method, thus for each pair of English-Khmer and French-Khmer, we employed the $10,000$ BPE tokens.

The model has parallel ASR and MT encoders. The ASR encoder uses 83-dimensional source speech features as the input, while the MT encoder takes another language text as the input where the vocabulary size is the input dimension. The decoder part is comprised of two cross-attentions. The summation operation was conducted to combine the 256-dimension of each attention and residual connection into a single 256-dimension output, as shown in Equation (3).

Table 1: *WER (%) of Khmer ASR on ECCC test set; ** and * indicates statistically significant difference with $p < 0.01$ and $p < 0.05$ from baseline, respectively.*

Model	WER (%) of the Khmer		
	Baseline	Joint$_{\mathrm{en}}$	Joint$_{\mathrm{fr}}$
w/o augmentation	23.6	22.2**	22.3**
w/ SpecAugment (SA)	22.2	21.1**	21.4**
w/ Speed perturbation (SP)	21.8	20.5**	20.6**
w/ SP + SA	21.4	**19.5****	20.2**

Table 2: *ASR improvement with proposed method for each group of speakers.*

Speaker Group	Hour	Average WER (%)		
		Baseline	Joint$_{\mathrm{en}}$	Relative
Witness	5	23.4	19.7**	15.8
Co-prosecutor	2	19.7	19.5	1.0
Civil-party	0.7	15.3	13.7**	10.5
Judge	0.3	17.0	17.1	-

Table 3: *ASR improvement with proposed method in accordance with baseline WER distribution.*

Baseline WER (%)	# utterance	Average WER (%)		
		Baseline	Joint$_{\mathrm{en}}$	Relative
$0-10$	$1,137$	4.5	5.3**	-
$10-20$	810	14.9	14.2*	4.7
$20-30$	538	25.8	23.6**	8.5
$30-40$	248	37.8	32.5**	14.0
$40-50$	165	49.4	43.3**	12.3
$50-100$	303	88.1	75.3**	14.5

Table 4: *ASR improvement with proposed method in accordance with MT BLEU distribution (English-to-Khmer).*

Baseline BLEU	# utterance	Average WER (%)		
		Baseline	Joint$_{\mathrm{en}}$	Relative
$0-10$	895	23.4	21.7**	7.3
$10-20$	$1,205$	20.2	18.4**	8.9
$20-30$	572	20.5	18.6**	9.3
$30-40$	268	18.7	17.1*	8.6
$40-50$	126	19.3	18.2	5.7
$50-100$	136	23.5	18.6**	20.9

4.3. System evaluation

The baseline MT of English to Khmer has a BLEU score of 14.44, which is better than the translation quality of French to Khmer, the BLEU score of which is 10.54. This is reasonable because English sentences were used as the source in sentence alignment and segmentation to Khmer and French as described in [9].

Table 1 presents the performance of our proposed method of joint training with English to Khmer MT (Joint$_{\mathrm{en}}$) and French to Khmer MT (Joint$_{\mathrm{fr}}$). The proposed method outperformed the baseline Khmer ASR model in all experimented models. All improvements are statistically significant ($p < 0.01$), but Joint$_{\mathrm{en}}$ gave a larger improvement compared to Joint$_{\mathrm{fr}}$. This is reasonable because English to Khmer MT has better performance. For the best performing model with SpecAugment (SA) and speed perturbation (SP), the proposed method reduced a

<table>
<tr><th>System</th><th>Output</th></tr>
<tr><td>Reference</td><td>នៅ |ថ្ងៃ |សាមសិប |ខែ |ដប់ |ពីរ |ឆ្នាំ |មួយពាន់ |ប្រាំបួនរយ |ចិតសិប |ប្រាំពីរ |ពេលនោះ |ខ្ញុំបាទ |កំពុង |តែ |ប្រមូល |ផល
<i>At |Day |30 |Month |10 |2 |Year |1000 |900 |70 |7 |That time |I |Was |Doing |Collect |Outcome</i></td></tr>
<tr><td>ASR</td><td>នៅ |ថ្ងៃ |សាមសិប |ខែ |ដប់ |ពីរ |ឆ្នាំ |មួយពាន់ |ប្រាំបួនរយ |ចិតសិប |ប្រាំពីរ |ពេលនោះ |ខ្ញុំបាទ |កំពុង |តែ |ប្រមូល |ផល |កសិកម្ម |នៅ
<i>At |Day |30 |Month |10 |2 |Year |1000 |900 |70 |7 |That time |I |Was |Doing |Collect |Outcome |Farming |At</i></td></tr>
<tr><td>MT</td><td>ថ្ងៃ |សាមសិប |ធ្នូ |មួយពាន់ |ប្រាំបួនរយ |ចិតសិប |ប្រាំពីរ |ដោយ |ពេល |ដែល |ខ្ញុំ |ទៅ |ធ្វើ |ស្រែ
<i>Day |30 |December |1000 |900 |70 |7 |By |When |That |I |Go |Do |Field</i></td></tr>
<tr><td>Joint_{en}</td><td>នៅ |ថ្ងៃ |សាមសិប |ខែ |ដប់ |ពីរ |ឆ្នាំ |មួយពាន់ |ប្រាំបួនរយ |ចិតសិប |ប្រាំពីរ |ពេលនោះ |ខ្ញុំបាទ |កំពុង |តែ |ប្រមូល |ផល |កសិកម្ម
<i>At |Day |30 |Month |10 |2 |Year |1000 |900 |70 |7 |That time |I |Was |Doing |Collect |Outcome |Farming</i></td></tr>
</table>

Figure 2: *Examples of the comparison of all methods in Khmer language, the italic text is the translated text into English.*

large margin of WER by 1.9% (8.9% relative).

Regarding the best result for Joint$_{\text{en}}$, Table 2 shows the system performance in each group of speakers. The proposed method had a significant improvement on "Witness" and "Civil-party," reducing the WER by 15.8% and 10.45% relative, respectively. These speaker groups include the victims of the Khmer Rouge regime, who are elderly and illiterate, thus had problems in their speech; they sometimes could not pronounce words correctly and exhibited disfluency and emotions in their speech during the trial. On the other hand, we did not obtain improvement for the group of "Judge" and "Co-prosecutor," who spoke fluently.

Table 3 presents the effectiveness of our proposed method in terms of the distribution of baseline WER. The worse the baseline ASR was, the more improvement is achieved with the proposed method. This trend is preferable in applications. In this case, the best improvement reduced the WER by 14.5% relative.

Table 4 presents the system performance in terms of the distribution of MT BLEU scores. It shows that a better MT performance generally results in a better improvement in the transcription of speech. This tendency is reasonable. With this result, the best MT BLEU score reduced the WER by 20.9% relative.

Figure 2 presents an example of output of the baseline ASR, MT, and the proposed method. We also investigated a possibility to combine the output hypotheses of ASR and MT. However, we found the hypotheses of MT is generally shorter (deletions of >30%) and much less accurate (substitutions of >30%) than the ASR hypotheses. This is because MT can have rephrasing without matching with speech (as annotated in "Blue text" in Figure 2) and less redundancy (no fillers, discourse markers). With this large difference between ASR and MT, we cannot combine the hypotheses of ASR and MT with ROVER. Moreover, it is not easy to combine two hypotheses with a simple voting mechanism. Instead, we propose a scheme to refer to MT for enhancing ASR hypotheses.

We also experimented the condition of replacing MT with ST, in which interpreters' speeches (e.g. English) are used for the input. In this setting, the WER was 20.0%, which is significantly improved the baseline but slightly lower than the originally proposed method using MT. This is due to the performance of the end-to-end ST. Since there is a limited number of interpreters in this corpus, separating ASR and MT is more practical.

4.4. Application to Fisher-CallHome-Spanish

To confirm that our proposed method can be generalized to other corpora, we conducted an experiment using Fisher-CallHome Spanish, which is a speech translation corpus of a conversational telephone speech in Spanish to English. It contains 160 hours of Spanish speech, corresponding transcription, and English translation text. The standard data preparation [23] was used, and the performances of Fisher-{dev, dev2, test} and CallHome-{devtest, evltest} were investigated.

Table 5: *WER (%) of speech transcription on Fisher-CallHome Spanish test set.*

Test set	w/ SP		w/ SP+SA	
	Baseline	Joint$_{\text{en}}$	Baseline	Joint$_{\text{en}}$
Fisher				
- dev	24.2	24.0	23.1	**22.8**
- dev2	23.6	23.1	22.5	**22.3**
- test	21.5	21.7	20.8	**20.5**
CallHome				
- devtest	41.1	40.5	40.2	**39.5**
- evltest	41.4	41.0	39.6	**39.4**

The network architecture of this implementation followed the given recipe in the ESPnet. Texts in English and Spanish were stripped of all punctuation and were lower-cased. The BPE was then used to tokenize the text by using $1,000$ tokens per language, which means that we employed $2,000$ BPE tokens in total.

Table 5 presents the results of the baseline ASR model and our proposed method (Joint$_{\text{en}}$) in each evaluation set. In all test sets, the joint training of Spanish ASR and English to Spanish MT improved the transcription of Spanish speech. Especially, with SA and SP data augmentations, Joint$_{\text{en}}$ reduced the WER up to 0.7% absolute (1.7% relative) in "devtest" of CallHome. These results demonstrate the generalization of the proposed method.

5. Conclusions

In this work, we have proposed a joint model of ASR and MT for improving the transcription of a low-resource language using a simultaneous translation from a rich-resource language. The proposed method was not only effective for improving the transcription in Khmer, but also in Spanish. The results demonstrate that translated knowledge is useful for enhancing the transcription of speech, especially for the lower-performance ASR with the higher translation quality of MT. This work is motivated from a language resource consideration, but in reality the proposed approach may be helpful in acoustically challenging conditions. Additionally, this method can be applied to many settings of simultaneous transcription and translation in multilingual meetings or court proceedings.

6. 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] 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 IEEE-ICASSP*, 2016.

[3] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in Joint CTC-Attention Based End-to-End Speech Recognition with a Deep CNN Encoder and RNN-LM," in *Proceedings of Interspeech*, 2017.

[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 NeurlPS*, 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. Ziemski, M. Junczys-Dowmunt, and B. Pouliquen, "The United Nations parallel corpus v1.0," in *Proceedings of LREC*, 2016, pp. 3530–3534.

[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 *Proceedings of IEEE-ICASSP*, 2020.

[8] D. Macháček, M. Žilinec, and O. Bojar, "Lost in Interpreting: Speech Translation from Source or Interpreter?" in *Proceedings of Interspeech*, 2021.

[9] 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.

[10] M. Paulik, S. Stuker, C. Fugen, T. Schultz, T. Schaaf, and A. Waibel, "Speech translation enhanced automatic speech recognition," in *Proceedings of IEEE-ASRU*, 2005.

[11] M. Paulik, C. Fügen, S. Stüker, T. Schultz, T. Schaaf, and A. Waibel, "Document driven machine translation enhanced asr," in *Proceedings of Eurospeech*, 2005.

[12] S. Khadivi, A. Zolnay, and H. Ney, "Automatic text dictation in computer-assisted translation," in *Proceedings of Eurospeech*, 2005.

[13] S. Khadivi, R. Zens, and H. Ney, "Integration of speech to computer-assisted translation using finite-state automata," in *Proceedings of COLING/ACL*, 2006.

[14] E. Vidal, F. Casacuberta, L. Rodríguez, J. Civera, and C. D. M. Hinarejos, "Computer-assisted translation using speech recognition," *IEEE TASLP*, 2006.

[15] S. Khadivi and H. Ney, "Integration of speech recognition and machine translation in computer-assisted translation," *IEEE TASLP*, 2008.

[16] A. Reddy and R. C. Rose, "Integration of statistical models for dictation of document translations in a machine-aided human translation task," *IEEE TASLP*, 2010.

[17] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proceedings of IEEE-ASRU*, 1997.

[18] P. Wang, T. N. Sainath, and R. J. Weiss, "Multitask Training with Text Data for End-to-End Speech Recognition," in *Proceedings of Interspeech*, 2021, pp. 2566–2570.

[19] B. Yusuf, A. Gandhe, and A. Sokolov, "USTED: Improving ASR with a Unified Speech and Text Encoder-Decoder," 2022.

[20] F. J. Och and H. Ney, "Statistical multi-source translation," in *Proceedings of MT Summit*, 2001.

[21] E. Garmash and C. Monz, "Ensemble learning for multi-source neural machine translation," in *Proceedings of COLING*, 2016, pp. 1409–1418.

[22] B. Zoph and K. Knight, "Multi-source neural translation," in *Proceedings of NAACL-HLT*, Jun. 2016, pp. 30–34.

[23] 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.

[24] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," 2020.

[25] 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*, 2019.

[26] 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.

[27] 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.

[28] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio Augmentation for Speech Recognition," in *Proceedings of Interspeech*, 2015.

[29] 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.

[30] R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in *Proceedings of ACL*, 2016.