IMPROVING END-TO-END SPEECH TRANSLATION MODEL WITH BERT-BASED CONTEXTUAL INFORMATION

Jeong-Uk Bang, Min-Kyu Lee, Seung Yun, Sang-Hun Kim

Electronics and Telecommunications Research Institute (ETRI), South Korea

ABSTRACT

This paper proposes an end-to-end speech translation system that utilizes contextual information. Contextual information helps clarify the meaning of the utterances. However, conventional end-to-end speech translation (E2E-ST) is primarily designed to handle single-utterance. Thus, we introduce a context encoder that extracts contextual information from previous translation results. Here, the context encoder obtains high-quality contextual information by adopting the BERT model. Then, we combine it with speech information extracted from speech signals to generate translation results. On the widely used TED-based speech translation corpus, we show that the results of the contextual E2E-ST model are significantly better than those of the single utterance-based E2E-ST model. Furthermore, we demonstrate that contextual information contributes to the processing of unclearly spoken utterances as well as ambiguity caused by pronouns and homophones.

Index Terms—Speech translation, contextual information, BERT, end-to-end models, Transformer

1. INTRODUCTION

Through metaverse and video conferencing platforms, we can easily meet each other in virtual space, even if we are physically far away in real life. However, conversing with users who speak different languages in the virtual space is still a challenge. Speech translation (ST) systems could overcome these language barriers by translating speech from one language into a sentence in another language. Conventional ST systems [1, 2] used automatic speech recognition (ASR) and machine translation (MT) models in a cascading manner (CAS-ST). Recently, end-to-end ST (E2E-ST) systems [3-6] in which two components are integrated into a single module have attracted attention.

Conversations between users in the virtual space can be easily logged and collected. In addition, the collected conversations can help clarify the meaning of a current conversation. In the field of MT, many methods [7-11] have been introduced to leverage the previous conversations. It has been reported that contextual information helps in handling homophones, pronouns, discourse connectives, and topic adaptation.

Contextual information is also helpful in E2E-ST. Most recently, Gaido *et al.* [12] showed that context-aware ST is less sensitive to segmentation errors that occurred in using voice activity detection. Also, Zhang *et al.* [13] demonstrated the effectiveness of context-aware ST. However, the training data for the E2E-ST model is very scarce [14], so they alone may not be enough to train the context-aware ST. Therefore, we need to utilize a pre-trained model with a large amount of MT corpus, such as the BERT model [15].

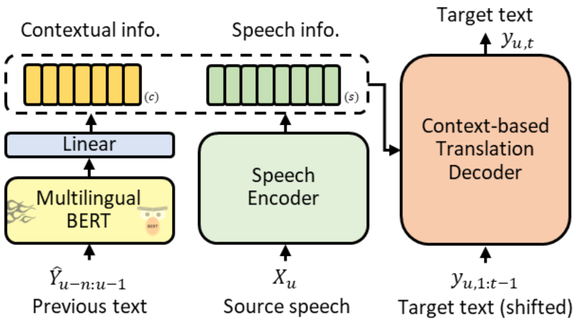

Fig. 1: A schematic illustration of the speech translation using BERT-based contextual information.

We propose an end-to-end ST model using the BERT to utilize contextual information. Our model consists of a context encoder that handles context information in previous sentences, a speech encoder that handles speech information in a current utterance, and a context-based translation decoder that performs translation based on both information. First, the context encoder takes the previous translation results as input and then extracts high-resolution word embedding vectors by using the pre-trained BERT model. The speech encoder extracts high-resolution speech embedding vectors, just like the existing single utterance-based ST model [3]. Finally, the translation decoder integrates the outputs of both encoders over a context-gate [7, 8] and then outputs a translated result. Our main contributions are the following: (i) We propose an end-to-end ST model with BERT-based contextual information. (ii) We show significant improvement over single utterance-based ST models on multiple ST datasets and demonstrate the effectiveness through ablation studies and generalizability. (iii) We show that contextual information contributes to the processing of unclearly spoken utterances.

2. CONTEXTUAL ST MODEL

Fig. 1 illustrates an end-to-end contextual ST model using BERT. Our goal is to find the most probable translated word sequence Y_u, when given a speech feature sequence X_u for the u-th utterance and a word sequence $Y_{u-n:u-1}$ for the n previous text together.

$$\begin{aligned}\hat{Y}_u &= \text{argmax}_{Y_u} p(Y_u|Y_{u-n:u-1}, X_u) \\ &= \text{argmax}_{Y_{u,1:T}} \textstyle\prod_{t=1}^{T} p\left(y_{u,t}|Y_{u-n:u-1}, X_u, y_{u,1:t-1}\right) \qquad (1)\end{aligned}$$

In the training phase, the previous sentences use ground-truth sentences (GT), not predicted by the ST model. On the other hand, in the inferencing phase, they use hypothesis sentences (HYP) obtained from the ST model with the previous utterances.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

2.1. Context encoder

Context encoder consists of the BERT model for obtaining contextual information and a linear layer for incorporating with the speech encoder. First, the BERT model takes a previous word sequence $Y_{u-n:u-1}$ as input. And then, it converts the word sequence into a token sequence via a BERT tokenizer $\mathrm{BertEnc}(\cdot)$ and computes embedding vectors for each token. Here, we use the publicly available multilingual BERT model [16], but our method is not specific to the BERT model. After that, the obtained embedding vectors are reduced to the same dimension as the speech encoder through a linear transformation $\mathrm{Linear}(\cdot)$. Finally, we obtain a context vector C via a layer normalization $\xi(\cdot)$.

$$C^{0}_{u-n:u-1} = \mathrm{BERT}(\mathrm{BertEnc}(Y_{u-n:u-1})) \tag{2}$$
$$C = \xi(\mathrm{Linear}(C^{0}_{u-n:u-1})) \tag{3}$$

2.2. Speech encoder

Speech encoder is the same architecture as that in a single utterance-based E2E-ST model [3, 17].

$$S^{0}_{u} = \mathrm{Linear}(\mathrm{Conv2D}(\mathrm{Conv2D}(X_u))) + \mathrm{PosEnc}(X_u) \tag{4}$$
$$\bar{S}^{n_s}_{u} = S^{n_s}_{u} + \mathrm{MHA}(\xi(S^{n_s-1}_{u})) \tag{5}$$
$$S^{n_s}_{u} = \bar{S}^{n_s}_{u} + \mathrm{FF}(\xi(\bar{S}^{n_s}_{u})) \tag{6}$$
$$S = \xi(S^{N_s}_{u}) \tag{7}$$

where, $\mathrm{Conv2D}(\cdot)$, $\mathrm{MHA}(\cdot)$, and $\mathrm{FF}(\cdot)$ represent a 2D convolution layer, multi-head attention and feed-forward network, respectively. n_s is the index of speech encoder block. For any unexplained notation, we refer to [18]. Finally, we compute a speech vector S through a layer normalization $\xi(\cdot)$ of the last block output $S^{N_s}_{u}$.

2.3. Context-based translation decoder

Translation decoder accepts previous output token sequence $y_{u,1:t-1}$, speech vector S, and context vector C. First, the input layer of translation decoder applies token embedding $\mathrm{TokEmb}(\cdot)$ and positional encoding $\mathrm{PosEnc}(\cdot)$.

$$D^{0}_{u,1:t-1} = \mathrm{TokEmb}(y_{u,1:t-1}) + \mathrm{PosEnc}(y_{u,1:t-1}) \tag{8}$$

Next, the decoder computes source and context attention vectors $SD^{n_d}_{u,1:t-1}$ and $CD^{n_d}_{u,1:t-1}$. Here, n_d denotes the index of translation decoder block and $\mathrm{MHA}(\cdot,\cdot,\cdot)$ takes the query, key, and value vector sequences [17].

$$\overline{D}^{n_d}_{u,1:t-1} = \xi(D^{n_d-1}_{u,1:t-1} + \mathrm{MHA}(\xi(D^{n_d-1}_{u,1:t-1}))) \tag{9}$$
$$SD^{n_d}_{u,1:t-1} = \xi(\overline{D}^{n_d-1}_{u,1:t-1} + \mathrm{MHA}(\xi(\overline{D}^{n_d-1}_{u,1:t-1}), S, S)) \tag{10}$$
$$CD^{n_d}_{u,1:t-1} = \xi(\overline{D}^{n_d-1}_{u,1:t-1} + \mathrm{MHA}(\xi(\overline{D}^{n_d-1}_{u,1:t-1}), C, C)) \tag{11}$$

To integrate two attention vectors, we use a context gate g_t [7]. It controls the flow from the speech encoder and context encoder. Then, the decoder computes a hidden vector $D^{n_d}_{u,1:t-1}$ with the following equations.

$$g_t = \sigma(W_{\mathrm{g}} \times \mathrm{Concat}(CD^{n_d}_{u,1:t-1}, SD^{n_d}_{u,1:t-1}) + b_g) \tag{12}$$
$$\overline{\overline{D}}^{n_d}_{u,1:t-1} = (1 - g_t) \odot CD^{n_d}_{u,1:t-1} + g_t \odot SD^{n_d}_{u,1:t-1} \tag{13}$$
$$D^{n_d}_{u,1:t-1} = \overline{\overline{D}}^{n_d}_{u,1:t-1} + \mathrm{FF}(\xi(\overline{\overline{D}}^{n_d}_{u,1:t-1})) \tag{14}$$

Finally, we obtain the probability of translated token $y_{u,t}$ by applying a layer normalization, a linear transformation, and a softmax layer $\mathrm{Softmax}(\cdot)$.

$$p\big(y_{u,t} \big| Y_{u-n:u-1}, X_u, y_{u,1:t-1}\big) = \mathrm{Softmax}(\mathrm{Linear}(\xi(D^{N_d}_{u,1:t-1}))) \tag{15}$$

For the contextual ST model training, we use a cross-entropy loss $L_u = -\log p(Y_u | Y_{u-n:u-1}, X_u)$ and for inferencing, we predict translated word sequence $\hat{Y}_u$ with the auto-regressive manner [17].

3. EXPERIMENTS

3.1. Datasets

We use two publicly available ST corpora, as shown in Table 1; EnKoST [19] for English-Korean (En-Ko) ST and MuST-C [14] for English-German (En-De), English-Spanish (En-Es), and English-French (En-Fr) ST. A dataset for each language pair consists of English audio files, English transcriptions, their translations, and their time information for each utterance. Each dataset offers standard training (Train), development (Dev), and evaluation (Eval) subsets. Each training set contains audio files of 408~559 hours with 234K~340K utterances. Each evaluation set contains audio files of 4.0~4.2 hours with 2502~2641 utterances. This paper mainly uses the EnKoST corpus to facilitate the validation of ST results.

Table 1. Statistic for each E2E-ST corpus.

Corpus	Lang (S-T)	#hours			#utts	#words (S/T)
		Train	Dev	Eval		
EnKoST [19]	En-Ko	559	2.5	4.0	340K	6.0M /4.2M
MuST-C [14]	En-De	408	2.5	4.1	234K	4.3M /4.0M
	En-Es	504	2.6	4.2	270K	5.3M /5.1M
	En-Fr	492	2.6	4.2	280K	5.2M /5.4M

3.2. Experiment setups

All experiments were performed by modifying the ESPnet toolkit [3]. In the context encoder, we use the multilingual BERT model (12-blocks, 768-hidden, 12-heads) and all model parameters are frozen during fine-tuning. The linear layer reduces the 768-hidden dimension obtained from BERT to the 256-hidden, the same as those used in the speech encoder.

In the speech encoder and context-based translation decoder, we use the Transformer encoder (12-blocks, 256-hidden, 4-heads) and decoder (6-blocks, 256-hidden, 4-heads). For speech features, we use 80-dimensional log-Mel filter-banks coefficients plus 3-dimensional pitch features per frame, resulting in 83-dimensional speech features per frame. The remaining parameters followed the MuST-C recipe of the ESPnet toolkit [3, 20].

The evaluation metrics are sacreBLEU [21] and METEOR [22]. For all language pairs, we use the default option of sacreBLEU, but we additionally use an external morphological analyzer [23] for English-Korean ST. We also measure the statistical significance of improvement with paired bootstrap resampling using sacreBLEU. All models are trained on four NVIDIA Titan RTX GPUs.

3.3. Baseline results

We first investigate the baseline results of single utterance-based models for each language pair. Table 2 shows word error rate (WER) and BLEU scores for ASR, MT, CAS-ST, and E2E-ST. From the MT results using English transcription as input, we can observe the translation difficulty of each language pair. In particular, despite using the largest amount of training data, the English-Korean MT model showed the lowest BLEU due to differences in linguistic characteristics. The CAS-ST model, which uses the ASR output as the MT input, outperforms the E2E-ST model, which translates English into other languages directly.

Table 2. Baseline results for ASR, MT, and ST. The ASR result indicates word error rate and the other results indicate BLEU.

Models	En-Ko	En-De	En-Es	En-Fr
ASR (↓)	9.8	13.2	12.4	12.5
MT (↑)	17.1	26.9	31.4	37.5
CAS-ST (↑)	16.1	22.6	27.3	31.5
E2E-ST (↑)	13.8	19.3	24.6	28.8

3.4. Contextual ST results

We investigate the contextual ST results. Table 3 compares the translation results of our contextual E2E-ST model with the single utterance-based E2E-ST model (baseline results in Table 2). In the tables, the "GT" experiment uses two ground-truth sentences as input of context encoder, and the "HYP" experiment uses the translation result of two previous utterances as context information.

In the English-Korean ST experiment, the ST model with GT outperforms baseline by +1.6 BLEU (significant at $p < 0.01$) and +1.2 METEOR. In addition, the ST model with HYP outperforms baseline by +1.0 BLEU (significant at $p < 0.01$) and +0.8 METEOR. As a result, the ST model using contextual information is much better than the baseline in all language pairs and shows the highest improvement of +2.0 BLEU (significant at $p < 0.01$) in the English-French ST experiment with GT.

Table 3. BLEU and METEOR scores for contextual E2E-ST. In the BLEU scores, "†" indicates a statistically significant difference ($p < 0.01$) from the baseline results (no context).

		En-Ko	En-De	En-Es	En-Fr
BLEU (↑)	**no context**	13.8	19.3	24.6	28.8
	w/ GT	**15.4**†	**20.8**†	**25.4**†	**30.8**†
	w/ HYP	14.8†	20.4†	25.2	29.5†
METEOR (↑)	**no context**	18.9	31.1	43.6	44.3
	w/ GT	**20.1**	**32.7**	**44.5**	**45.5**
	w/ HYP	19.7	32.2	44.1	44.7

3.5. Number of context sentences

Table 4 shows the performance according to the number of context sentences in the English-Korean ST experiment. In the table, the ST model with two context sentences ('2' in Table 4) shows the best performance. Moreover, all contextual ST models outperform the single utterance-based ST model ('0' in Table 4) regardless of the number of context sentences.

On the other hand, the performance degrades with three context sentences ('3' in Table 4) than two context sentences. When using more than three context sentences, the contextual ST model has difficulty in obtaining useful information for ST from them. This is because Transformer models often suffer when given too long of an input sequence [13], and this degradation has been reported equally in the document-level MT [8]. Therefore, we use two sentences as contextual information for the following experiments.

Table 4. BLEU and METEOR scores according to the number of context sentences. In the BLEU scores, "†" indicates a statistically significant difference ($p < 0.01$) from the baseline (#Context: 0).

#Context		**0**	**1**	**2**	**3**
BLEU (↑)	**GT**	13.8	14.4†	**15.4**†	14.9†
	HYP		14.0	**14.8**†	14.4†
METEOR (↑)	**GT**	18.9	19.3	**20.1**	19.8
	HYP		19.2	**19.7**	19.4

4. ANALYSIS

4.1. Ablation studies

To quantify the effect of BERT and context gate, we conduct an ablation study by removing one of them from our model. Table 5 shows the results of the ablation study. We first remove the pre-trained BERT (Table 5(a)) from the context encoder. That is, in the context encoder, Equation 2 for the BERT model is replaced by Equation 8 for the token embedding. The BERT model contributes to an improvement of +0.8 BLEU (significant at $p < 0.01$), showing the importance of using the pre-trained model.

We also employ a Transformer encoder instead of the BERT (Table 5(b)). Here, we used a single Transformer block, and using more blocks did not improve performance. The BERT-based context encoder shows a slight improvement of +0.2 BLEU in the HYP experiment and a significant improvement of +0.6 BLEU (significant at $p < 0.01$) in the GT experiment. As a result, the BERT-based context encoder has the potential to further improve performance when given higher quality context sentences.

Table 5. Ablation study results. Numbers indicate BLEU and METEOR scores when using ground-truth (left) and hypothesis (right) sentences as context. In the BLEU scores, "†" indicates a statistically significant difference ($p < 0.01$) from our model.

Models	BLEU (↑)	METEOR (↑)
Our model	**15.4 / 14.8**	**20.1 / 19.7**
(a) w/o pre-trained BERT (linear layer)	14.6† / 14.1†	19.6 / 19.3
(b) w/o pre-trained BERT (Transformer encoder)	14.8† / 14.6	19.8 / 19.6
(c) w/o context gating (averaging)	14.4† / 14.2†	19.4 / 19.2

If we do not employ context gating (Table 5(c)), the average of the speech vector and the context vector is fed to the translation decoder. It shows that the context gating contributes a gain of +1.0 BLEU (significant at $p < 0.01$) in the GT experiment. The reason is that context gating regulates information between speech encoder and context encoder. As a result, we confirmed that the pre-trained BERT and context gate contributed significantly to the performance of the contextual ST model.

Table 8: Comparison of translation results between baseline and our model. Two examples have different contextual sentences with the same single utterance. Italics indicate examples translated into English.

	Example A	**Example B**
Previous sentences	한 곳은 미국입니다. (*One* ***place*** *is the United States.*)	한 명은 미국인입니다. (*One* ***person*** *is the American.*)
Source speech	One is South Korea.	
Baseline results	한 명은 한국입니다. (*One* ***person*** *is in South Korea.*)	
Our results	한 곳은 한국입니다. (*One* ***place*** *is South Korea.*)	한 명은 한국인입니다. (*One* ***person*** *is South Korean.*)

4.2. Generalizability

Table 6 shows the comparison results according to contextual information quality. This suggests the lower and upper bounds of our model. When a random sentence (Table 6(b)) not related to the current utterance is given as contextual information, the contextual ST model shows 13.7 BLEU, which is similar to the 13.8 BLEU (Table 6(a)) of the single utterance-based ST model. It is because our model can decide whether or not to use given contextual information via context gating. When the previous sentence (Table 6(c)) is given as contextual information, our ST model significantly outperforms the single utterance-based ST model as described above. Moreover, although unlikely to occur, when the target sentence (Table 6(d)) is given as contextual information, our ST model shows 18.6 BLEU, outperforming 17.1 BLEU (Table 2) of the MT model.

Table 6. Comparison results of the BLEU and METEOR scores according to contextual information quality.

Context type	BLEU (↑)	METEOR (↑)
(a) No context	13.8	18.9
(b) Random sentence	13.7	18.7
(c) Previous sentence	15.4 (14.8)	20.1 (19.7)
(d) Target sentence	18.6	22.1

4.3. Results according to WER

We tested the effect of contextual information for unclearly spoken utterances. We divided the evaluation set on the English-Korean ST corpus into two groups on the basis of the average WER, 9.8 %, of speech recognition in Table 2. Then, we compared the improvement in BLEU score of the two groups. As a result, the contextual information contributes more to unclearly spoken utterances with higher WER in speech recognition, as shown in Table 7.

The previous studies [12, 13] on context-aware ST have focused on solving speech segmentation errors that occur when using VAD and analyzing whether it has similar effects to document-level MT. On the other hand, we demonstrated that contextual information could contribute to the processing of unclearly spoken utterances.

Table 7. Comparison results of the BLEU score improvement effect by word error rate (WER).

WER (%)	Baseline	Proposed	Diff
< 9.8	15.2	16.5	+1.3
≥ 9.8	11.6	13.6	+2.0

4.4. Comparison of translation results

Table 8 shows the translation results of the single utterance-based ST model and the contextual ST model. Given the source speech, the single utterance-based ST model produced the fixed results regardless of the previous sentence. On the other hand, the contextual ST model produced different translation results ("place" in Example A and "person" in Example B) depending on the previous sentence.

Fig. 2 shows an attention map between the context encoder and the translation decoder for Example A in Table 8. In Fig. 2, the circles indicate that the "place" in the output focuses on the "place" in the previous sentence. This example shows that the proposed model works correctly for our purposes.

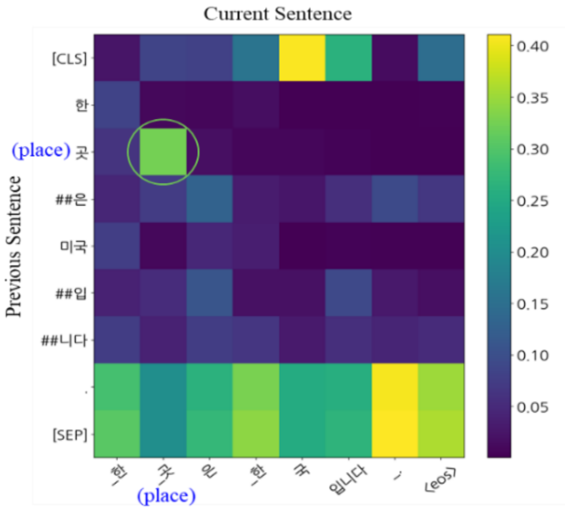

Fig. 2: An attention map for Example A. The x-axis and y-axis correspond to the current sentence and the previous sentence, respectively.

5. CONCLUSION

We proposed an end-to-end contextual ST model with the pre-trained BERT model. The proposed model significantly outperforms a single utterance-based speech translation model on multiple speech translation datasets. We also show the effect of the pre-trained model from the ablation study and present the upper and lower bounds for our model according to the contextual information quality. We demonstrate that contextual information can contribute not only to the processing of ambiguity caused by pronouns but also to the processing of unclearly spoken utterances.

6. ACKNOWLEDGEMENTS

This work was supported by the Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [22ZS1100, Core Technology Research for Self-Improving Integrated Artificial Intelligence System].

7. REFERENCES

[1] H. Ney, "Speech translation: Coupling of recognition and translation," In *Proceedings of ICASSP*, IEEE, 1999, pp. 517-520.

[2] E. Cho, J. Niehues, and A. Waibel, "NMT-Based Segmentation and Punctuation Insertion for Real-Time Spoken Language Translation," In *Proceedings of Interspeech*, 2017, pp. 2645-2649.

[3] H. Inaguma, S. Kiyono, K. Duh, S. Karita, N. E. Y. Soplin, T. Hayashi, and S. Watanabe, "ESPnet-ST: All-in-one speech translation toolkit," In *Proceedings of ACL*, 2020, pp. 302-311.

[4] A. Bérard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," In *Proceedings of NIPS workshop on end-to-end learning for speech and audio processing*, 2016.

[5] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, and J. Pino, "Fairseq S2T: Fast Speech-to-Text Modeling with Fairseq," In *Proceedings of AACL-IJCNLP*, 2020, pp. 33-39.

[6] Y. Jia, M. Johnson, W. Macherey, R. J. Weiss, Y. Cao, C. C. Chiu, et al. "Leveraging weakly supervised data to improve end-to-end speech-to-text translation," In *Proceedings of ICASSP*, IEEE, 2019, pp. 7180-7184.

[7] L. M. Werlen, D. Ram, N. Pappas, and J. Henderson, "Document-Level Neural Machine Translation with Hierarchical Attention Networks," In *Proceedings of EMNLP*, 2018, pp. 2947-2954.

[8] Y. Kim, D. T. Tran, and H. Ney, "When and Why is Document-level Context Useful in Neural Machine Translation?" In *Proceedings of DiscoMT*, 2019, pp. 24-34.

[9] L. Wang, Z. Tu, A. Way, and Q. Liu, "Exploiting Cross-Sentence Context for Neural Machine Translation," In *Proceedings of the EMNLP*, 2017, pp. 2826-2831.

[10] J. Zhang, H. Luan, M. Sun, F. Zhai, J. Xu, M. Zhang, and Y. Liu, "Improving the Transformer Translation Model with Document-Level Context," In *Proceedings of EMNLP*, 2018, pp. 533-542.

[11] R. Bawden, R. Sennrich, A. Birch, and B. Haddow, "Evaluating Discourse Phenomena in Neural Machine Translation," In *Proceedings of NAACL-HLT*, 2018, pp. 1304-1313.

[12] M. Gaido, M. A. Di Gangi, M. Negri, M. Cettolo, and M. Turchi, "Contextualized Translation of Automatically Segmented Speech," In *Proceedings of Interspeech* 2020, pp. 1471-1475.

[13] B. Zhang, I. Titov, B. Haddow, and R. Sennrich, "Beyond Sentence-Level End-to-End Speech Translation: Context Helps," In *Proceedings of ACL-IJCNLP*, 2021, pp. 2566-2578.

[14] R. Cattoni, M. A. Di Gangi, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: A multilingual corpus for end-to-end speech translation," *Computer Speech & Language*, 66, pp. 101155.

[15] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," In *Proceedings of NAACL-HLP*, 2019, pp. 4171-4186.

[16] "Multilingual BERT model," https://huggingface.co/bert-base-multilingual-cased, [Online; accessed 2021-07-06].

[17] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, iang, Z., et al. "A comparative study on transformer vs rnn in speech applications," In *Proceedings of ASRU*, IEEE, 2019, pp. 449-456.

[18] T. Hori, N. Moritz, C. Hori, and J. Le Roux, "Transformer-Based Long-Context End-to-End Speech Recognition," In *Proceedings of Interspeech*, 2020, pp. 5011-5015.

[19] "EnKoST dataset for English-Korean Speech Translation," https://nanum.etri.re.kr/share/seungyun/EnKoSTCv10, [Online; accessed 2021-07-06].

[20] "MuST-C speech translation recipe for ESPnet toolkit," https://github.com/espnet/espnet/tree/master/egs/must_c/st1, [Online; accessed 2021-07-06].

[21] M. Post, "A Call for Clarity in Reporting BLEU Scores," In *Proceedings of ACL*, 2018 pp. 186-191.

[22] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," In *Proceedings of ACL*, 2005, pp. 65-72.

[23] E. L. Park, and S. Cho, "KoNLPy: Korean natural language processing in Python. In *Proceedings of HLT*, 2014, pp. 133-136.