Interspeech 2022
18-22 September 2022, Incheon, Korea

Speech Segmentation Optimization using Segmented Bilingual Speech Corpus for End-to-end Speech Translation

Ryo Fukuda, Katsuhito Sudoh, Satoshi Nakamura

Nara Institute of Science and Technology, Japan
{fukuda.ryo.fo3, sudoh, s-nakamura}@is.naist.jp

Abstract

Speech segmentation, which splits long speech into short segments, is essential for speech translation (ST). Popular VAD tools like WebRTC VAD[1] have generally relied on pause-based segmentation. Unfortunately, pauses in speech do not necessarily match sentence boundaries, and sentences can be connected by a very short pause that is difficult to detect by VAD. In this study, we propose a speech segmentation method using a binary classification model trained using a segmented bilingual speech corpus. We also propose a hybrid method that combines VAD and the above speech segmentation method. Experimental results reveal that the proposed method is more suitable for cascade and end-to-end ST systems than conventional segmentation methods. The hybrid approach further improves the translation performance.

Index Terms: speech translation, segmentation, bilingual speech corpus

1. Introduction

The segmentation of continuous speech is one process required for speech translation (ST). In text-to-text machine translation (MT), an input text is usually segmented into sentences using punctuation marks as boundaries. However, such explicit boundaries are unavailable in ST. Pre-segmented speech in a bilingual speech corpus can be used when training an ST model, but not in realistic scenarios. Since existing ST systems cannot directly translate long continuous speech, automatic speech segmentation is needed.

Pause-based segmentation using voice activity detection (VAD) is commonly used for segmenting speech for preprocessing automatic speech recognition (ASR) and ST, even though pauses do not necessarily coincide with semantic boundaries. Over-segmentation, in which a silence interval fragments a sentence, and under-segmentation, in which multiple sentences are included in one segment while ignoring a short pause, are problems that reduce the performances of ASR and ST [1].

In this study, we propose a novel speech segmentation method[2] based on the segment boundaries in a bilingual speech corpus. ST corpora usually include bilingual text segment pairs with corresponding source language speech segments. If we can similarly segment input speech to such ST corpora, we might successfully bridge the gap between training and inference. Our proposed method uses a Transformer encoder [2] that was trained to predict segment boundaries as a frame-level sequence labeling task for speech inputs.

The proposed method is applicable to cascade and end-to-end STs because it directly segments speech. Cascade ST is a traditional ST system that consists of an ASR model and a text-to-text MT model. An end-to-end ST uses a single model to directly translate source language speech into target language text. We limit our scope to a conventional end-to-end ST, which treats segmentation as an independent process from ST. Integrating a segmentation function into end-to-end ST is future work.

We conducted experiments with cascade and end-to-end STs on MuST-C [3] for English-German and English-Japanese. In the English-German experiments, the proposed method achieved improvements of -9.6 WER and 3.2 BLEU for cascade ST and 2.7 BLEU for end-to-end ST compared to pause-based segmentation. In the English-Japanese experiments, the improvements were -9.6 WER and 0.5 BLEU for cascade ST and 0.6 BLEU for end-to-end ST. A hybrid method with the proposed segmentation model and VAD further improved both language pairs.

2. Related work

Early studies on segmentation for ST considered modeling with the Markov decision process [4, 5], conditional random fields [6, 7], and support vector machines [8, 9, 10, 11]. They focused on cascade ST systems that consist of an ASR model and a statistical machine translation (SMT) model, which were superseded by newer ST systems based on neural machine translation (NMT).

In recent studies, many speech segmentation methods based on VAD have been proposed for ST. Gaido et al. [12] and Inaguma et al. [13] used the heuristic concatenation of VAD segments up to a fixed length to address the over-segmentation problem. Gállego et al. [14] used a pre-trained ASR model called wav2vec 2.0 [15] for silence detection. Yoshimura et al. [16] used an RNN-based ASR model to consider consecutive blank symbols ("_") as a segment boundary in decoding using connectionist temporal classification (CTC). Such CTC-based speech segmentation has an advantage; it is easier to intuitively control segment lengths than with a conventional VAD because the number of consecutive blank symbols that is considered a boundary can be adjusted as hyperparameters. However, these methods often split audio at inappropriate boundaries for ST because they mainly segment speech based on long pauses.

Re-segmentation using ASR transcripts is widely used in cascade STs. Improvements in MT performance have been reported by re-segmenting transcriptions to sentence units using punctuation restoration [7, 11, 17, 18, 19] and language models [20, 21]. Unfortunately, they are difficult to use in end-to-end ST and cannot prevent ASR errors due to pause-based segmentation. We discuss ASR degradation due to speech segmentation with VAD in section 5.1.

Finally, we refer to speech segmentation methods based on segmented speech corpora that are more relevant to our study.

[1] https://github.com/wiseman/py-webrtcvad

[2] There is an independent work by Tsiamas et al., (2022) (https://arxiv.org/abs/2202.04774).

Copyright © 2022 ISCA

10.21437/Interspeech.2022-11382

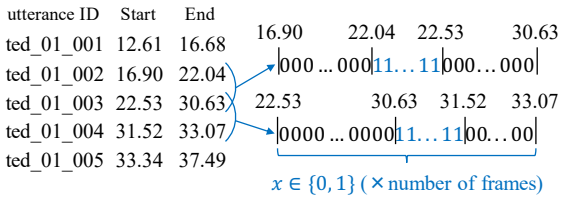

Figure 1: *Data extraction examples from MuST-C: Labels 0 and 1 are assigned to each FBANK frame based on each segment's starting and ending times.*

Wan et al. [1] introduced a re-segmentation model for modifying the segment boundaries of ASR output using movie and TV subtitle corpora. Wang et al. [22] and Iranzo-Sánchez et al. [23] proposed an RNN-based text segmentation model using the segment boundaries of a bilingual speech corpus. Unlike these methods that require ASR transcripts, our proposed method directly segments speech and can be used in an end-to-end ST as well as in a cascade ST.

3. Proposed method

Our method defines speech segmentation as a frame-level sequence labeling task of acoustic features. This section describes the following details of it: the process that extracts training data for the speech segmentation task from the speech translation corpus (3.1), the architecture of the proposed model (3.2), and the training (3.3) and inference algorithms (3.4).

3.1. Data extraction

We use a bilingual speech corpus that includes speech segments aligned to sentence-like unit text as training data for speech segmentation from continuous speech into units suitable for translation. Data extraction examples are shown in Fig. 1. Two consecutive segments are concatenated and assigned a label, $x \in \{0, 1\}$, representing that the corresponding frame is inside (0) and outside (1) of the utterance.

3.2. Segmentation model

The architecture of the proposed speech segmentation model is illustrated in Fig. 2. It consists of a 2D convolution layer, Transformer encoder layers, and an output layer (Linear+Softmax), which outputs label probability $\hat{x}_n \in R^2$ at the n-th frame of the convolution layer. The convolution layer reduces the sequence length by a quarter to handle long speech sequences. Downsampling is applied to teacher label x to align its length with the model's output. Here, labels are simply extracted at regular time intervals calculated from the input-output length ratio.

3.3. Training

The model is trained using the data extracted in section 3.1 and learns to minimize cross-entropy loss $\mathcal{L}_{seg}(\hat{x}, x)$ between prediction $\hat{x}$ and label x:

$$\mathcal{L}_{seg}(\hat{x}, x) := -\sum_{n=1}^{N} \left\{ w_s \log \frac{\exp(\hat{x}_{n,1})}{\exp(\hat{x}_{n,0} + \hat{x}_{n,1})} x_{n,1} + (1 - w_s) \log \frac{\exp(\hat{x}_{n,0})}{\exp(\hat{x}_{n,0} + \hat{x}_{n,1})} x_{n,0} \right\}, \quad (1)$$

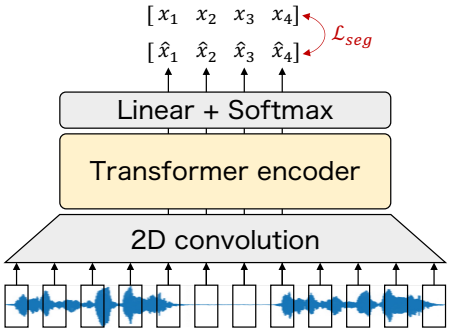

Figure 2: *Transformer encoder-based speech segmentation model*

where w_s is a hyperparameter that adjusts the weights of the unbalanced labels. Since most of the labels are 0 (within utterances), we increase the weight on the loss of label 1 (outside utterances). We set $w_s = 0.9$ based on the best loss of the validation data in a preliminary experiment.

3.4. Inference

During inference, speech is segmented at a fixed-length T and input independently into the segmentation model. Fixed-length segments are then resegmented according to the labels predicted by the segmentation model. The segmentation model simply selects label $l_n \in \{0, 1\}$ with the highest probability at each time n:

$$l_n := \arg\max(x_n). \quad (2)$$

3.4.1. Hybrid method

In addition, to make the predictions more appropriate, we combined the model predictions calculated in Eq. (2) with the results of the VAD tool:

$$l_n := \begin{cases} \arg\max(x_n) \wedge \mathrm{vad}_n & (len(n-1) < maxlen) \\ \arg\max(x_n) \vee \mathrm{vad}_n & (len(n-1) \geq maxlen), \end{cases} \quad (3)$$

where $\mathrm{vad}_n \in \{0, 1\}$ is the VAD output at time n, corresponding to the active and inactive frames. Segment length $len(n)$ at time n is denoted as follows:

$$len(n) := \begin{cases} len(n-1) + 1 & (l_n = 1) \\ 0 & (l_n = 0) \end{cases}, \quad (4)$$

where $len(0) := 0$. Eq. (3) implies that a frame can only be included in a segment boundary when segmentation model **and** VAD agree. This agreement constraint may result in very long segments, so we relax it not to require the agreement once the segment length exceeds $maxlen$. In that case, a frame can be included in a segment boundary if either the segmentation model **or** the VAD allows it.

4. Experimental setup

4.1. Tasks

We conducted experiments with English to German and English to Japanese ST. We used MuST-C v1 for the English-German and v2 for the English-Japanese experiments. Each dataset consisted of triplets of segmented English speech, transcripts, and target language translations. The English-German and English-Japanese datasets contained about 230k and 330k segments. As

Table 1: *Model settings: †version 0.10.3*

Settings (ESPnet† options)	ASR	ST	MT
Epochs (epochs)	45	100	
Encoder layers (elayers)	12		6
Decoder layers (elayers)	6		
FNN dimensions (eunits, dunits)	2048		
Attention dimensions (adim)	256		
Attention heads (aheads)	4		
Mini-batch (batch-size)	64		96
Gradient accumulation (accum-grad)	2		1
Gradient clipping (grad-clip)	5		
Learning rate (transformer-lr)	5	2.5	1
Warmup (transformer-warmup-steps)	25000		
Label smoothing (lsm-weight)	0.1		
Attention dropout (dropout-rate)	0.1		

acoustic features, we used 83-dimensional vectors consisting of an 80-dimensional log Mel filterbank (FBANK) extracted by Kaldi[3] and 3-dimensional pitch information. We preprocessed the text data with Byte Pair Encoding (BPE) to split the sentences into subwords with SentencePiece [24]. A dictionary with a vocabulary of 8,000 words was shared between the source and target languages.

To evaluate the performance, we aligned the outputs of a model for automatically segmented speeches by each segmentation method (section 4.3) to the reference text with an edit distance-based algorithm [25]. We then calculated WER for ASR and BLEU for MT and ST.

4.2. ST systems

We used the Transformer implementation of ESPnet[4] [26, 27] to build the ASR and MT models for the cascade ST and an ST model for the end-to-end ST. For the ASR and ST models using acoustic features as input, we added a 2D-convolution layer before the transformer encoder layers. The model settings are shown in Table 1. We trained the ASR model with hybrid CTC/attention [28] that incorporates the CTC loss into the Transformer. The weight on the CTC loss was set to 0.3. The ST model parameters were initialized with the encoder of the ASR model. The parameters of each model were stored for each epoch. After training to the maximum number of epochs, we averaged the parameters of five epochs with the highest validation scores for evaluation.

4.3. Segmentation methods

4.3.1. Baseline 1: VAD

As a baseline, we used WebRTC VAD, a GMM-based VAD. We tried nine conditions in the range of *Frame size*={10, 20, 30 ms} and *Aggressiveness*={1, 2, 3}.

4.3.2. Baseline 2: fixed-length

Fixed-length is a length-based approach that splits speech at a pre-defined fixed length [5]. Although over- and under-segmentations are likely to occur because the method does not take acoustic and linguistic clues into account, its advantage is that the segment length was kept constant. We tried ten fixed-length settings with parameters ranging from 4 to 40 seconds at 4-second intervals.

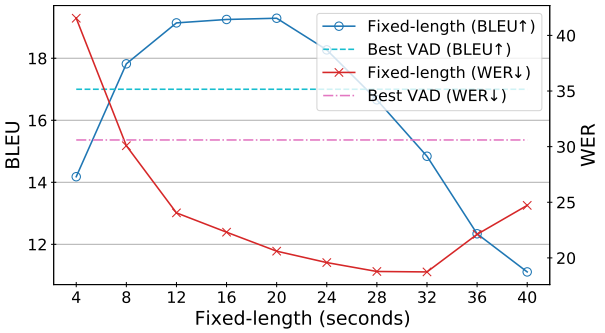

Figure 3: *Baseline scores on English-German cascade ST: Vertical axis indicates BLEU (left) and WER scores (right) for ASR and MT models.*

4.3.3. Speech segmentation optimization

We performed speech segmentation using our proposed model described in section 3.2. During training, two consecutive segments were concatenated, as described in section 3.1. The model's setting was almost identical as that of the ASR encoder shown in Table 1. However, since the speech segment concatenation roughly doubles the average length of the input, we set the number of mini-batches to 32 and the number of gradient accumulations to 4. Fixed-length T of the input during the inference was set to 20 seconds based on the best score of the validation data.

4.3.4. Hybrid method

We also performed the hybrid method with our model and VAD, as introduced in section 3.4.1. For the hybrid method, we used a WebRTC VAD with (frame size, aggressiveness) = (10 ms, 2) and set *maxlen* to ten seconds.

5. Results and discussion

5.1. Baselines

Figure 3 shows the ASR and MT results by cascade ST for each fixed-length. For comparison, the scores for the setting with the highest BLEU of VAD are shown as straight lines (Best VAD). With the fixed-length approach, WER and BLEU improved in proportion to the input length and deteriorated after reaching a certain length. This result suggests that longer segments can prevent the degradation of the ASR and ST performances due to automatic segmentation. On the other hand, we identified an upper limit to the segment length that the model successfully handled, which depends on its capability. Therefore, automatic speech segmentation is important to prevent ST performance degradation. In addition, the best VAD results were worse than the best fixed-length results (Best Fixed-length)[5], suggesting that over- and under-segmentation due to pause-based segmentation significantly reduced the ASR and ST performances.

[3]https://github.com/kaldi-asr/kaldi
[4]https://github.com/espnet/espnet

[5]A similar trend was shown in a previous study [12].

Table 2: *Results measured in WER and BLEU for tst-COMMON on MuST-C English-German cascade and end-to-end ST.*

	Cascade ST		End-to-end ST
	WER	BLEU	BLEU
Oracle	12.60	23.59	22.50
Best VAD	30.59	17.02	16.40
Best Fixed-length	20.60	19.29	17.96
Our model	20.99	20.18	19.10
+VAD hybrid	**19.06**	**20.99**	**19.87**

Table 3: *Results measured in WER and BLEU for tst-COMMON on MuST-C English-Japanese cascade and end-to-end ST.*

	Cascade ST		End-to-end ST
	WER	BLEU	BLEU
Oracle	9.30	12.50	10.60
Best VAD	25.81	9.26	8.14
Best Fixed-length	18.89	9.64	8.52
Our model	16.21	9.71	8.77
+VAD hybrid	**13.67**	**10.60**	**9.24**

5.2. Proposed method

Table 2 shows the overall results of the English-German experiments. Our model outperformed VAD and the fixed-length segmentations for both cascade and end-to-end STs. Roughly speaking, there are improvements in +3 BLEU for the Best VAD and in +1 BLEU for the Best Fixed-length. This suggests that our model can split the speech into segments that correspond to sentence-like units suitable for translation.

In addition, the hybrid method with VAD significantly improved both cascade (-1.93 WER and +0.81 BLEU) and end-to-end STs (+0.77 BLEU). We confirmed that a hybrid method with our model and VAD greatly improved the translation performance. However, room for improvement remains compared to the oracle segments contained by the MuST-C corpus.

Table 3 shows the overall results of the English-Japanese experiments. They resemble those in English-German; our model outperformed the existing methods, and a hybrid with VAD achieved more significant improvements. We conclude that the proposed method is also effective for distant language pairs.

5.3. Case study

Table 4 shows an example of the ASR and MT outputs of the cascade ST using VAD and our segmentation model. In VAD, more over- and under-segmentation occurred compared to the oracle segments. In the example, the best VAD resulted in over-segmentation ("*bonobos are* ■ *together*") and under-segmentation ("*relative that ...*"). These errors caused differences in the oracle segments in the MT results. On the other hand, our model split the speech at a boundary close to an oracle segment and obtained the same translation results.

Figure 4 shows an example of the waveforms and their segmentation positions. The top two waveforms show that our model and VAD caused over-segmentation. As the bottom waveform shows, hybrid decoding split the audio at a boundary near the oracle and alleviated the problem by requiring an agreement between our model and the VAD.

Table 4: *Example of ASR and MT outputs with segmentation positions:* ■ *indicates segment boundaries.*

Oracle (ASR)	*bonobos are together with chimpanzees you aposre living closest relative* ■
Oracle (MT)	*Bonobos sind zusammen mit Schimpansen, Sie leben am nachsten Verwandten.* ■
Best VAD (ASR)	*bonobos are* ■ *together with chimpanzees you aposre living closest relative that ...*
Best VAD (MT)	*Bonobos sind es.* ■ *Zusammen mit Schimpansen leben Sie im Verhaltnis zum ...*
Our model (ASR)	*bonobos are together with chimpanzees you aposre living closest relative* ■
Our model (MT)	*Bonobos sind zusammen mit Schimpansen, Sie leben am nachsten Verwandten.* ■

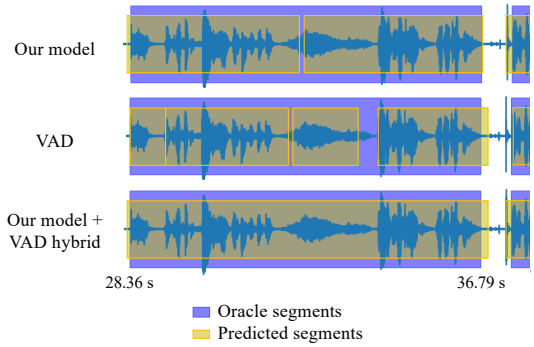

Figure 4: *Visualization of waveforms and segmentation positions*

6. Conclusions

We proposed a speech segmentation method based on a bilingual speech corpus. Our method directly split speech into segments that correspond to sentence-like units to bridge the gap between training and inference. Our experimental results showed the effectiveness of the proposed method compared to conventional segmentation methods on both cascade and end-to-end STs. We also demonstrated that combining the predictions of our model and VAD further improved the translation performance.

Future work will integrate a segmentation function into an end-to-end ST. In this study, we treated segmentation as an independent process from ST, although it should be included in the *end-to-end* ST. We will investigate ways to integrate our proposed segmentation method into ST for streaming or online processing. Future work will also further investigate different domains, other language pairs, and noisy environments and improve the segmentation model to reduce the need for VAD.

7. Acknowledgements

Part of this work was supported by JST SPRING Grant Number JPMJSP2140 and JSPS KAKENHI Grant Numbers JP21H05054 and JP21H03500.

8. References

[1] D. Wan, C. Kedzie, F. Ladhak, E. Turcan, P. Galuščáková, E. Zotkina, Z. P. Jiang, P. Bell, and K. McKeown, "Segmenting subtitles for correcting asr segmentation errors," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 2842–2854.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[3] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "Must-c: a multilingual speech translation corpus," in *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2019, pp. 2012–2017.

[4] S. Mansour, "Morphtagger: Hmm-based arabic segmentation for statistical machine translation," in *Proceedings of the 7th International Workshop on Spoken Language Translation: Papers*, 2010.

[5] M. Sinclair, P. Bell, A. Birch, and F. McInnes, "A semi-markov model for speech segmentation with an utterance-break prior," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[6] T. Nguyen and S. Vogel, "Context-based Arabic morphological analysis for machine translation," in *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Manchester, England: Coling 2008 Organizing Committee, Aug. 2008, pp. 135–142. [Online]. Available: https://aclanthology.org/W08-2118

[7] W. Lu and H. T. Ng, "Better punctuation prediction with dynamic conditional random fields," in *Proceedings of the 2010 conference on empirical methods in natural language processing*, 2010, pp. 177–186.

[8] M. Diab, K. Hacioglu, and D. Jurafsky, "Automatic tagging of Arabic text: From raw text to base phrase chunks," in *Proceedings of HLT-NAACL 2004: Short Papers*. Boston, Massachusetts, USA: Association for Computational Linguistics, May 2 - May 7 2004, pp. 149–152. [Online]. Available: https://aclanthology.org/N04-4038

[9] F. Sadat and N. Habash, "Combination of Arabic preprocessing schemes for statistical machine translation," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia: Association for Computational Linguistics, Jul. 2006, pp. 1–8. [Online]. Available: https://aclanthology.org/P06-1001

[10] E. Matusov, D. Hillard, M. Magimai-Doss, D. Hakkani-Tur, M. Ostendorf, and H. Ney, "Improving speech translation with automatic boundary prediction," in *Proceedings of Interspeech 2007*, 2007, pp. 2449–2452.

[11] V. K. Rangarajan Sridhar, J. Chen, S. Bangalore, A. Ljolje, and R. Chengalvarayan, "Segmentation strategies for streaming speech translation," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 230–238. [Online]. Available: https://aclanthology.org/N13-1023

[12] M. Gaido, M. Negri, M. Cettolo, and M. Turchi, "Beyond voice activity detection: Hybrid audio segmentation for direct speech translation," *CoRR*, vol. abs/2104.11710, 2021. [Online]. Available: https://arxiv.org/abs/2104.11710

[13] H. Inaguma, B. Yan, S. Dalmia, P. Guo, J. Shi, K. Duh, and S. Watanabe, "ESPnet-ST IWSLT 2021 offline speech translation system," in *Proceedings of the 18th International Conference on Spoken Language Translation*. Bangkok, Thailand (online): Association for Computational Linguistics, Aug. 2021, pp. 100–109. [Online]. Available: https://aclanthology.org/2021.iwslt-1.10

[14] G. I. Gállego, I. Tsiamas, C. Escolano, J. A. Fonollosa, and M. R. Costa-jussà, "End-to-end speech translation with pre-trained models and adapters: Upc at iwslt 2021," in *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, 2021, pp. 110–119.

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

[16] T. Yoshimura, T. Hayashi, K. Takeda, and S. Watanabe, "End-to-end automatic speech recognition integrated with ctc-based voice activity detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6999–7003.

[17] E. Cho, J. Niehues, K. Kilgour, and A. Waibel, "Punctuation insertion for real-time spoken language translation," in *Proceedings of the Eleventh International Workshop on Spoken Language Translation*, 2015.

[18] T.-L. Ha, J. Niehues, E. Cho, M. Mediani, and A. Waibel, "The kit translation systems for iwslt 2015," in *Proceedings of the Eleventh International Workshop on Spoken Language Translation*, 2015.

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

[20] A. Stolcke and E. Shriberg, "Automatic linguistic segmentation of conversational speech," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 2. IEEE, 1996, pp. 1005–1008.

[21] X. Wang, A. Finch, M. Utiyama, and E. Sumita, "An efficient and effective online sentence segmenter for simultaneous interpretation," in *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, 2016, pp. 139–148.

[22] X. Wang, M. Utiyama, and E. Sumita, "Online sentence segmentation for simultaneous interpretation using multi-shifted recurrent neural network," in *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, 2019, pp. 1–11.

[23] J. Iranzo-Sánchez, A. Giménez Pastor, J. A. Silvestre-Cerdà, P. Baquero-Arnal, J. Civera Saiz, and A. Juan, "Direct segmentation models for streaming speech translation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 2599–2611. [Online]. Available: https://aclanthology.org/2020.emnlp-main.206

[24] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *EMNLP (Demonstration)*, 2018.

[25] E. Matusov, G. Leusch, O. Bender, and H. Ney, "Evaluating machine translation output with automatic sentence segmentation," in *Proceedings of the Second International Workshop on Spoken Language Translation*, 2005.

[26] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," in *Proceedings of Interspeech 2018*, 2018, pp. 2207–2211.

[27] 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 the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, Jul. 2020, pp. 302–311. [Online]. Available: https://aclanthology.org/2020.acl-demos.34

[28] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.