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 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 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 ("\n") 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.
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, \( x \in \{0, 1\} \) (× number of frames), representing that the corresponding frame is inside (0) or 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 \( \hat{x}_{n, 0} \in \mathbb{R}^2 \) 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 \( \mathcal{L}_{\text{seq}}(\hat{x}, x) \) between prediction \( \hat{x} \) and label \( x \):

\[
\mathcal{L}_{\text{seq}}(\hat{x}, x) := - \sum_{n=1}^{N} \left[ w_n \log \frac{\exp(\hat{x}_{n, 1})}{\exp(\hat{x}_{n, 0}) + \exp(\hat{x}_{n, 1})} x_{n, 1} + (1 - w_n) \log \frac{\exp(\hat{x}_{n, 0})}{\exp(\hat{x}_{n, 0}) + \exp(\hat{x}_{n, 1})} x_{n, 0} \right],
\]

where \( w_n \) 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 \( w_n = 0.9 \) 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 \( n \):

\[
l_n := \text{arg max}(x_n).
\]

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:

\[
l_n := \begin{cases} 
\text{arg max}(x_n) \land \text{vad}_n & (\text{len}(n-1) < \text{maxlen}) \\
\text{arg max}(x_n) \lor \text{vad}_n & (\text{len}(n-1) \geq \text{maxlen})
\end{cases},
\]

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

\[
\text{len}(n) := \begin{cases} 
\text{len}(n-1) + 1 & (l_n = 1) \\
0 & (l_n = 0)
\end{cases},
\]

where \( \text{len}(0) := 0 \). 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 \( \text{maxlen} \). 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 v1.1 for the English-German and v2 for the English-Japanese experiments. Each dataset consisted of triplets of segmented English speech, transcriptions, 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

<table>
<thead>
<tr>
<th>Settings (ESPnet† options)</th>
<th>ASR</th>
<th>ST</th>
<th>MT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epochs (epochs)</td>
<td>45</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Encoder layers (elayers)</td>
<td>12</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Decoder layers (elayers)</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FNN dimensions (eunits, dunits)</td>
<td>2048</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attention dimensions (adim)</td>
<td>256</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attention heads (ahead)</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mini-batch (batch-size)</td>
<td>64</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>Gradient accumulation (accum-grad)</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Gradient clipping (grad-clip)</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learning rate (transformer-lr)</td>
<td>25000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Label smoothing (lsm-weight)</td>
<td>0.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attention dropout (dropout-rate)</td>
<td>0.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Acoustic features, we used 83-dimensional vectors consisting of an 80-dimensional log Mel filterbank (FBANK) extracted by Kaldi2 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† [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.

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.

\(^5\)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.

<table>
<thead>
<tr>
<th></th>
<th>Cascade ST</th>
<th>End-to-end ST</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WER</td>
<td>BLEU</td>
</tr>
<tr>
<td>Oracle</td>
<td>12.60</td>
<td>23.59</td>
</tr>
<tr>
<td>Best VAD</td>
<td>30.59</td>
<td>17.02</td>
</tr>
<tr>
<td>Best Fixed-length</td>
<td>20.60</td>
<td>19.29</td>
</tr>
<tr>
<td>Our model</td>
<td>20.99</td>
<td>19.18</td>
</tr>
<tr>
<td>+VAD hybrid</td>
<td>19.06</td>
<td>20.99</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Cascade ST</th>
<th>End-to-end ST</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WER</td>
<td>BLEU</td>
</tr>
<tr>
<td>Oracle</td>
<td>9.30</td>
<td>12.50</td>
</tr>
<tr>
<td>Best VAD</td>
<td>25.81</td>
<td>9.26</td>
</tr>
<tr>
<td>Best Fixed-length</td>
<td>18.89</td>
<td>9.64</td>
</tr>
<tr>
<td>Our model</td>
<td>16.21</td>
<td>9.71</td>
</tr>
<tr>
<td>+VAD hybrid</td>
<td>13.67</td>
<td>10.60</td>
</tr>
</tbody>
</table>

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 with chimpanzees you pose living closest relative") 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 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.

<table>
<thead>
<tr>
<th></th>
<th>bonobos are together with chimpanzees you pose living closest relative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle (MT)</td>
<td>Bonobos sind zusammen mit Schimpansen, Sie lebten am nachsten Verwandten.</td>
</tr>
<tr>
<td>Best VAD (ASR)</td>
<td>bonobos sind zusammen mit Chimpanzees you are living closest relative...</td>
</tr>
<tr>
<td>Best VAD (MT)</td>
<td>Bonobos sind zusammen mit Schimpansen leben Sie am Verhältnis zu...</td>
</tr>
<tr>
<td>Our model (ASR)</td>
<td>Bonobos sind zusammen mit Schimpansen, Sie lebten am nachsten Verwandten.</td>
</tr>
<tr>
<td>Our model (MT)</td>
<td>Bonobos sind zusammen mit Schimpansen, Sie lebten am nachsten Verwandten.</td>
</tr>
</tbody>
</table>

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


