

ORIGINAL ARTICLE

English–Korean speech translation corpus (EnKoST-C): Construction procedure and evaluation results

Jeong-Uk Bang¹  | Joon-Gyu Maeng² | Jun Park¹ | Seung Yun¹ | Sang-Hun Kim¹

¹Integrated Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea

²ICT-Computer Software, University of Science and Technology, Daejeon, Republic of Korea

Correspondence

Seung Yun, Integrated Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea.
Email: syun@etri.re.kr

Funding information

Electronics and Telecommunications Research Institute, Grant/Award Number: 22ZS1100

Abstract

We present an English–Korean speech translation corpus, named EnKoST-C. End-to-end model training for speech translation tasks often suffers from a lack of parallel data, such as speech data in the source language and equivalent text data in the target language. Most available public speech translation corpora were developed for European languages, and there is currently no public corpus for English–Korean end-to-end speech translation. Thus, we created an EnKoST-C centered on TED Talks. In this process, we enhance the sentence alignment approach using the subtitle time information and bilingual sentence embedding information. As a result, we built a 559-h English–Korean speech translation corpus. The proposed sentence alignment approach showed excellent performance of 0.96 f-measure score. We also show the baseline performance of an English–Korean speech translation model trained with EnKoST-C. The EnKoST-C is freely available on a Korean government open data hub site.

KEYWORDS

bilingual sentence alignment, corpus, end-to-end speech translation, speech-to-text translation, spoken language translation

1 | INTRODUCTION

Speech translation (ST) systems are indispensable to overcome language barriers. Generally, ST technology generates a sentence in a target language from a speech in a source language [1]. Conventional ST tasks use automatic speech recognition (ASR) and machine translation (MT) components to process speech in a cascading manner [2]. In this paper, we denote this process as CAS-ST. Although both components have shown improvement in

recent years, it suffers from a large computational cost and latency because each of them is an independent module. Recently, an end-to-end ST (E2E-ST) task [1] in which two components were integrated into a single module has attracted attention. Using the E2E-ST approach, we can expect a simplified system architecture, latency reduction, and mitigation of error propagation between ASR and MT [3].

The data required for the CAS-ST process have been accumulated over a long time [4–6]. However, training data for E2E-ST models, which are composed of speech data in the source language and the corresponding text in

Jeong-Uk Bang and Joon-Gyu Maeng contributed equally to this work.

This is an Open Access article distributed under the term of Korea Open Government License (KOGL) Type 4: Source Indication + Commercial Use Prohibition + Change Prohibition (<http://www.kogl.or.kr/info/licenseTypeEn.do>).

1225-6463/\$ © 2022 ETRI

the target language, are scarce. Additionally, most available E2E-ST data are European language oriented [7–9], and, to the best of our knowledge, no data for English–Korean E2E-ST models exist to date.

TED Talks, which provide English audio files and subtitles translated into various languages, including Korean, can be a good source to build an ST corpus. The multilingual ST corpus (MuST-C) proposed by Cattoni and others [10] is a representative ST corpus based on TED Talks [11] that is widely used as a benchmark corpus for E2E-ST systems [12,13]. The MuST-C was constructed using a procedure based on automatic alignment. Additionally, they provided detailed construction methods that can be extended to other languages.

This paper introduces a TED-based English–Korean ST corpus (EnKoST-C). This corpus was constructed using an automatic collection method based on sentence alignment [10]. First, we collected English audio files, English subtitles, and Korean subtitles from the TED website [11]. Then, we performed sentence-level bilingual text alignment to build an English–Korean parallel corpus. Here, we proposed a novel alignment method using both subtitle time information and a similarity measure based on bilingual sentence embeddings. This method resolves the alignment problems that occur due to the character differences between English and Korean. Next, we identified speech segments by speech-to-text alignment between the English audio files and the English sentences. Finally, we removed potentially noisy speech segments by measuring the average word duration and sentence length. As a result, we obtained a 559-h EnKoST-C from 3138 TED Talks.

We compared the proposed bilingual sentence alignment method with previous methods [14–17] and present the baseline performance of E2E-ST models trained with EnKoST-C. To facilitate the expansion to other languages, this EnKoST-C is designed with the same data partition as MuST-C. Thus, the baseline performance can be easily reproduced using the MuST-C recipe in the ESPnet toolkit [18,19].

Our primary contributions are as follows. (1) We introduce an EnKoST-C for the first time. (2) We propose a novel sentence alignment method that uses time information and a similarity measure based on bilingual sentence embeddings. (3) We present baseline performance and example scripts. Please note that EnKoST-C is released under the Creative Commons license, Attribution-Noncommercial-No Derivatives (CC BY-NC-ND) 4.0 international and is freely available at <https://nanum.etri.re.kr/share/seungyun/EnKoSTCv10>.

The rest of this paper is organized as follows. Existing ST corpora are described in Section 2. The proposed corpus creation procedure is presented in Section 3.

Sections 4 and 5 present the proposed text alignment results and the baseline performance of the E2E-ST models trained using the EnKoST-C. Conclusion is presented in Section 6.

2 | RELEATED WORK

ST corpora comprise pairs of source language speech segments and text translated into the target language. In the following sections, existing corpus creation methodologies and bilingual text alignment methods are described.

2.1 | Speech translation corpora

Benchmark corpora frequently used for ST tasks are listed in Table 1. Most available ST corpora were built for European languages, which have similar linguistic characteristics. There are a few ST corpora for Asian languages, such as Chinese (Zh), Japanese (Ja), and Vietnamese (Vi). Most recently, the Korean–English ST corpus called Kosp2e [20] has been released. However, there is no EnKoST-C, which consists of English speech and Korean translation.

Parallel data for an ST corpus can be collected from professional translators, by crowdsourcing, or via an automated process based on sentence alignment. In Table 1, CoVoST 2 [23] and MLST [22] were built by professional translators, and How2 [21] and Fisher-CallHome Spanish–English [9] were built by crowdsourcing [24,25]. Even though building a large

TABLE 1 Benchmark corpora for speech translation

Corpus	Languages	Hours
MuST-C [10]	En → {Ar, Cs, De, Es, Fa, Fr, It, Nl, Pt, Ro, Ru, Tr, Vi, Zh}	237–504
CoVoST 2 [21]	En → {De, Tr, Fa, Sv, Mn, Zh, Cy, Ca, Sl, Et, Id, Ar, Ta, Lv, Ja}	364
How2 [21]	En → Pt	300
LibriVoxDeEn [8]	De → En	110
Libri-Trans [7]	En → Fr	236
MSLT [22]	En ↔ {Fr, De, Ja, Zh}	2–5
Fisher-CallHome Spanish [9]	Es → En	170
Kosp2e [20]	Ko → En	198

Notes: The languages are denoted by their ISO codes. CoVoST 2 can handle translations from 21 languages into English and from English into 15 languages.

corpus is difficult due to the labor-intensive and computational costs, both methods can build a high-quality corpus. However, MuST-C [10], LibriVoxDeEn [8], and Libri-Trans [7] were built via automatic alignment techniques. With such techniques, bilingual alignment is performed at the text level between the source and target texts. Then, speech-to-text alignment is performed between the source text and corresponding audio segments [10].

TED Talks are ideal for creating an ST corpus automatically because they provide good quality speech recordings with human-curated transcriptions and translations in multiple languages. MuST-C [10] collected TED data and then built an MuST-C using the Gargantua toolkit [16] for bilingual text alignment and the Gentle toolkit [26] for a speech-to-text alignment. This method can be adopted as the procedure for building the EnKoST-C.

2.2 | Bilingual text alignment methods

In this section, we briefly describe four approaches for bilingual text alignment, that is, length-, translation-, timestamp-, and embedding-based approaches. The length-based approach [15,27] compares the length of the source and target sentences. The translation-based approach [15,16] relies on lexical correspondences between the target sentence and the translated source sentence. Bilingual text alignment can be performed quickly with the subtitle timestamp approach [14,28] by relying on a temporal indexing of subtitles lines. Vecalign [17] is a representative example of the sentence embedding approach [29]. Vecalign uses a dynamic program to compare sentence embeddings to align source and target language sentences. Among these four bilingual text

alignment approaches, sentence embedding provides state-of-the-art accuracy.

3 | PROPOSED METHOD

The overall process of the proposed corpus creation method using data from TED Talks is shown in Figure 1. The creation process involves data collection, data preprocessing, bilingual text alignment, speech-to-text alignment, and data filtering. The proposed method differs from the MuST-C creation method [10] in the bilingual text alignment and data filtering steps.

3.1 | Data collection

Initially, data collection involves downloading available videos (MPEG4) and their English and Korean subtitles (VTT) from the TED website. The collected data comprise 575.4 h of 3401 talks, which is approximately 81.0% of all available TED Talks as of September 2020.

3.2 | Data preprocessing

The collected audio and text data require preprocessing. To extract audio files (RIFF WAV) from the video, we use the FFmpeg toolkit (sampling rate, 16 kHz; sample size, 16-bit) [30].

Because better results are obtained when bilingual text alignment is done at the sentence level [14], preprocessing the text data involves converting subtitle-level texts to sentence-level texts and estimating the timestamps. It should be noted that the subtitle boundary is inconsistent with the sentence boundary. Additionally,

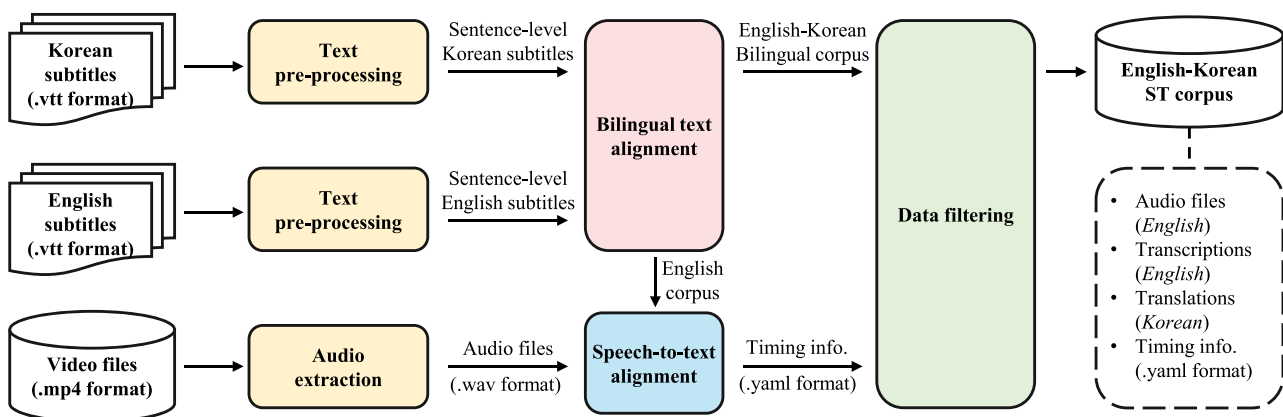


FIGURE 1 Overall process of the proposed corpus creation method

the start and end times of each sentence must be estimated because we use the time information in the bilingual text alignment.

To obtain the start and end timestamps of each sentence, the subtitle concatenation and splitting approach proposed by Tiedemann [28] are followed. First, we concatenate consecutive subtitles that complete the sentence. Here, sentence boundaries are determined by strong punctuation marks, that is, full stop (.), exclamation (!), and question (?) marks. When subtitles are concatenated, the timestamps of the merged subtitles are obtained from the corresponding original subtitles.

Then, we split the concatenated subtitle into sentences based on the punctuation marks. We approximate the start and end timestamps of each sentence within a single subtitle according to the following equations:

$$d(S) = e(S) - s(S), \quad (1)$$

$$e(S_{\text{sent}}) = s(S_{\text{sent}}) + d(S_{\text{subt}}) \times \frac{l(S_{\text{sent}})}{l(S_{\text{subt}})}, \quad (2)$$

where $s(S)$ and $e(S)$ are the start and end times of subtitle S , respectively; S_{sent} is a sentence within a single subtitle S_{subt} ; and $d(S)$ and $l(S)$ denote the duration and character length of subtitle S , respectively. Equation (2) estimates the end time $e(S_{\text{sent}})$ for a sentence S_{sent} within a single subtitle S_{subt} . Here, $s(S_{\text{sent}})$ is assigned as the start time of the subtitle S_{subt} or the end time of the previous sentence.

3.3 | Bilingual text alignment

We use time information and a similarity measure based on sentence embeddings to perform bilingual sentence alignment. In English and Korean sentences, we match sentence pairs with a similar timestamp. Then, unaligned portions are rearranged by finding pairs with the highest cosine similarity between the bilingual sentence embeddings.

Initially, alignment is based on the start time and duration between the source sentences $S_{1:|s|} = \{S_1, \dots, S_{|s|}\}$ and target sentences $T_{1:|t|} = \{T_1, \dots, T_{|t|}\}$, as shown in Algorithm 1. For each source sentence S_i , we select the target sentence T_j that has the closest start time. Then, they are considered a matched pair $match(S_i, T_j)$, if S_i and T_j satisfy the following two conditions:

$$\text{condition 1: } |s(S_i) - s(T_j)| < \delta, \quad (3)$$

$$\text{condition 2: } |d(S_i) - d(T_j)| < \delta, \quad (4)$$

Algorithm 1 Bilingual Text Alignment Algorithm

Input: Two sentences with timestamps $S_{1:|s|}$, $T_{1:|t|}$

Output: Two aligned sentences $\hat{S}_{1:|a|}$, $\hat{T}_{1:|a|}$

```

1: // '←' denotes 'assign' and '⇐' denotes 'append'
2: function align( $S_{1:|s|}$ ,  $T_{1:|t|}$ )
3:   // Stage 1: timestamp-based alignment
4:   for each sentence in  $S_{1:|s|}$  do
5:      $T_j \leftarrow$  the closest sentence to  $S_i$ 
6:     if match( $S_{i:i+1}, T_j$ ) then  $M \leftarrow (S_{i:i+1}, T_j)$ 
7:     else if match( $S_i, T_{j:j+1}$ ) then  $M \leftarrow (S_i, T_{j:j+1})$ 
8:     if match( $S_i, T_j$ ) then  $M \leftarrow (S_i, T_j)$ 
9:      $\hat{S}_{TS}, \hat{T}_{TS} \leftarrow M$ 
10:  end for
11:
12:  // Stage 2: sentence embedding-based alignment
13:   $C \leftarrow$  get_unaligned_chunks( $S$ ,  $T$ ,  $\hat{S}_{TS}$ ,  $\hat{T}_{TS}$ )
14:  for each chunk in  $C_{1:|c|}$  do
15:     $V \leftarrow$  run_vecalign( $C_k$ )
16:     $\hat{S}_{EM}, \hat{T}_{EM} \leftarrow$  remove_insertion_deletion( $V$ )
17:  end for
18:
19:   $\hat{S}_{1:|a|}, \hat{T}_{1:|a|} \leftarrow (\hat{S}_{TS}, \hat{T}_{TS}) \cup (\hat{S}_{EM}, \hat{T}_{EM})$ 
20:  return ( $\hat{S}_{1:|a|}$ ,  $\hat{T}_{1:|a|}$ )
21: end function

```

where $s(S)$ and $d(S)$ denote the start time and duration of subtitle S , respectively, whereas δ is the predefined threshold. The smaller the value, the more relevant sentence pairs can be aligned. Moreover, we allow 2-to-1 and 1-to-2 alignments, denoted by $match(S_{i:i+1}, T_j)$ and $match(S_i, T_{j:j+1})$, respectively. In this paper, $S_{i:i+1}$ and $T_{j:j+1}$ concatenate consecutive sentences. As a result, we obtain the aligned pairs of the source and target sentence, \hat{S}_{TS} and \hat{T}_{TS} , respectively.

Next, we perform alignment based on the sentence embeddings for the unaligned sentences that remain after timestamp-based alignment. We refer to these unaligned portions as chunks C . By partitioning these chunks, we essentially divide a large text alignment problem into several smaller problems of the same type. Then, unaligned sentences of each chunk C_k are aligned using Vecalign [17], which match sentence pairs with the highest similarity cosine between the bilingual sentence embeddings. The alignment results V obtained from the Vecalign include insertion and deletion sentences that do not exist in a parallel sentence and many-to-many alignment sentences. Therefore, we remove insertion and deletion sentences that are unnecessary for building the ST corpus. Then, the preserved pairs of source and target sentences are stored in \hat{S}_{EM} and \hat{T}_{EM} , respectively.

Finally, we combine the \hat{S}_{TS} and \hat{T}_{TS} obtained using the time information and the \hat{S}_{EM} and \hat{T}_{EM} obtained

using the sentence embeddings. As a result, we output the bilingual corpus consisting of English transcription $\hat{S}_{1:|a|}$ and Korean translation $\hat{T}_{1:|a|}$.

3.4 | Speech-to-text alignment

We perform speech-to-text alignment between the English audio files and English text sentences to find speech segments. Here, the sentence is the English transcription obtained in the bilingual text alignment. Like other corpora [7,10], we use Gentle [26], which is an off-the-shelf English forced aligner based on the Kaldi ASR toolkit [31]. We then create a YAML file containing time information, that is, the start time and duration, for each segment.

We also remove the segments where the start or end time could not be found and record the number of sentences removed in each talk and the number of unaligned words in each sentence. This alignment information will be used to weed out potentially noise segments in the data filtering step.

3.5 | Data filtering

Here, noise segments of unsuccessful forced alignments and nonspeech events are removed. Initially, we follow two noise segment filtering criteria used in MuST-C [10]. The first rule discards entire talks if the proportion of unaligned words is greater than or equal to 15% of the total words. The second rule removes all sentences having unaligned words. By applying these filtering rules, 263 talks are removed.

However, some segments still contained misaligned portions, primarily in nonspeech segments. Here, the nonspeech segments can be recognized incorrectly because they are not the target covered by the Gentle forced aligner. Therefore, we adopted average word duration (AWD) [32] as an additional filter. An AWD is the ratio between the duration and number of words in the segment. Figure 2 illustrates the segment distribution and cumulative distribution based on the AWD value.

The left- and rightmost portions of Figure 2 represent segments that are uttered abnormally: quickly or slowly. These outlier portions are representative of music, laughter, and applause, which are unnecessary for ST model training. Therefore, we select a training set within the range $0.15 < \text{AWD} < 0.65$. This range was determined considering that the TED Talks utterances are slightly faster than broadcast speech [32]. Given this range, we rejected 2.9% of the 575.4 h of TED Talks leading to 558.8 h of training data.

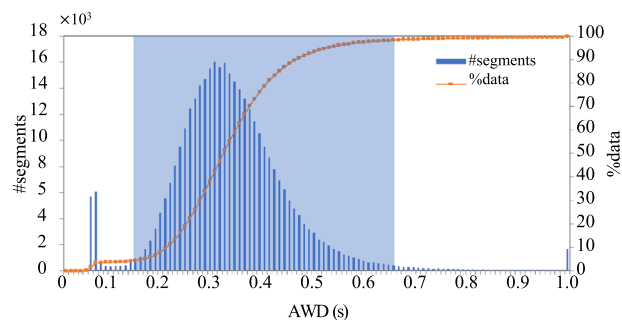


FIGURE 2 Segment distribution and cumulative duration in percentage data based on the AWD. The shaded portion corresponds to the region where $0.15 < \text{AWD} < 0.65$

4 | BILINGUAL TEXT ALIGNMENT RESULTS

4.1 | Evaluation metrics

To evaluate our bilingual text alignment method, we selected 49 TED Talks, which include the development and test sets used by MuST-C [10]. We created reference alignments $S_{1:|r|}$ and $T_{1:|r|}$ by manually aligning the sentences of all sentence pairs of English–Korean subtitles from these talks. We then obtained results $\hat{S}_{1:|a|}$ and $\hat{T}_{1:|a|}$ using the proposed bilingual sentence alignment method. We compared our alignment results with their references using an F_1 score, which is the harmonic mean of the precision and recall [14]. Here, the precision is the number of correct alignments divided by the number of proposed alignments, including those not aligned correctly, whereas recall is the number of correct alignments divided by the number of reference alignments.

4.2 | Sentence alignment results

We compared the performance of the proposed method with four off-the-shelf bilingual sentence aligners: Alignerzilla [14], Hunalign [15], Gargantua [16], and Vecalign [17]. All experiments were performed on the same test datasets: dev, tst-COMMON, and tst-HE. Table 2 shows the alignment approaches used in each aligner, the number of sentence pairs output from each aligner and their precision, recall, and F_1 scores. To demonstrate the proposed alignment method, we first compared our timestamp-based alignment results. In Table 2, the “Proposed A” row shows the alignment performance using only the timestamp-based approach. This approach differs from “Alignerzilla A” in that it performs strict alignment by including *condition 1* (3). As a result, the number of aligned sentence pairs decreased from 5836 to

TABLE 2 Comparison of sentence alignment quality

Algorithm	Approach	Aligned sentence pairs	dev			tst-COMMON			tst-HE			Average		
			P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Alignerzilla A [14]	TS_A	5836	0.79	0.83	0.81	0.84	0.88	0.86	0.85	0.89	0.87	0.83	0.87	0.85
Alignerzilla B [14]	TS_B + LN + DT	5663	0.71	0.74	0.73	0.78	0.82	0.81	0.79	0.82	0.80	0.77	0.80	0.78
Hunalign A [15]	LN	6148	0.19	0.18	0.19	0.37	0.35	0.36	0.24	0.23	0.23	0.29	0.28	0.28
Hunalign B [15]	LN + DT	6133	0.46	0.46	0.46	0.62	0.61	0.62	0.58	0.58	0.58	0.56	0.56	0.56
Gargantua A [16]	WP + LN + MT	5231	0.62	0.58	0.60	0.53	0.80	0.82	0.66	0.57	0.61	0.74	0.69	0.71
Gargantua B [16]	MP + LN + MT	5598	0.66	0.67	0.66	0.85	0.86	0.86	0.69	0.68	0.69	0.77	0.76	0.76
Vecalgn [17]	SE	5673	0.90	0.92	0.91	0.95	0.96	0.95	0.92	0.93	0.92	0.93	0.94	0.94
Proposed A	TS_A	4877	0.95	0.83	0.89	0.98	0.85	0.91	0.99	0.86	0.92	0.97	0.85	0.91
Proposed B	TS_B	4381	0.97	0.74	0.84	0.99	0.77	0.87	0.99	0.81	0.89	0.99	0.77	0.87
Proposed C	TS_B + SE	5611	0.94	0.94	0.94	0.97	0.97	0.97	0.96	0.96	0.96	0.96	0.96	0.96

Notes: TS_A and TS_B denote the timestamp-based approach with thresholds of 0.95 and 0.475, respectively. LN denotes a length-based approach. DT and MT denote dictionary-based and machine translation-based approaches, respectively. SE denotes a sentence embedding-based approach. WP and MP perform word-level and morpheme-level sentence alignment, respectively. Default aligner parameters were used.

4877; however, the precision and F_1 scores improved by 14 and 6 points, respectively. Additionally, as shown for “Proposed B,” when the threshold δ was reduced from 0.95 to 0.475, the F_1 score decreased by approximately four points, but the precision increased by approximately two points. Here, precision is a more important measure than F_1 , because we subsequently perform realignment on unaligned portions.

Furthermore, the length- and translation-based alignment results, that is, “Alignerzilla B,” “Hunalign A and B,” and “Gargantua A and B,” returned poor alignment quality for English–Korean sentence alignment. This is because Korean differs significantly from English in terms of basic units, word-spacing rules, and sentence word order. Moreover, like MuST-C [10], we tried alignment by tokenizing Korean subtitle text; however, the alignment results were still poor, as shown in “Gargantua B.” In Table 2, “Vecalgn,” which only used the similarity measure based on bilingual sentence embeddings, outperformed other off-the-shelf aligners in our test datasets.

Our approach, that is, “Proposed C,” outperformed all other compared alignment approaches. Although the exclusively timestamp-based approach, that is, “Proposed B,” returned higher precision values than “Vecalgn,” it outputs a smaller number of alignment results. The proposed approach, which uses the time information and

similarity measure based on bilingual sentence embedding, generated more alignment results than “Proposed A and B.” Additionally, the F_1 score for the proposed approach was the highest (0.96). These results demonstrate that the timestamp and sentence embedding-based approaches are complementary.

We also compared the impact of threshold δ in (2) and (3) on the timestamp-based approach. Table 3 shows the alignment quality according to various threshold values. In the table, we observed that a threshold with 0.5 shows the highest F_1 score, and the timestamp-based aligning results cover 79.5% of overall aligning results. As the threshold increased, more timestamp-based aligning was performed, but the average F_1 score decreased. For this reason, we adopted a threshold of 0.475, equal to the default setting in the Alignerzilla [14], as the final value of the proposed method.

TED Talks always have well-written subtitles and accurate timestamps. Thus, the proposed method can be used effectively to create our EnKoST-C.

4.3 | Error analysis

We performed an error analysis on the 49 talks used in the alignment experiment. Out of the 5611 sentence pairs

TABLE 3 Alignment quality according to various threshold values

Threshold (δ)	0.0	0.5	1.0	3.0	5.0
#Aligned sents	5672	5612	5596	5590	5602
TS stage (%)	0 (0.0)	4460 (79.5)	4958 (88.6)	5309 (95.0)	5474 (97.9)
EM stage (%)	5672 (100.0)	1152 (20.5)	638 (11.4)	281 (5.0)	128 (2.3)
F_1 score	0.94	0.96	0.95	0.87	0.81

Note: TS denotes the number of sentences obtained from the timestamp-based alignment stage, and EM denotes the number of sentences obtained from the sentence embedding-based alignment stage.

obtained as our alignment results, we extracted 235 misaligned sentence pairs. Among them, deletion and insertion errors, which are partial alignment with missing and additional information, accounted for the most at 47.1% and 38.5%, respectively. On the other hand, the substitution error, which is wrong alignments, was relatively small at 2.3%. The rest were alignment errors that occurred in nonspeech notation, additional explanations, and repeated sentences. We observed that most of the alignment errors occurred in short sentences. Table 4 shows some examples of alignment errors.

5 | CORPUS CONSTRUCTION RESULTS

5.1 | Corpus structure

Our corpus, EnKoST-C, contains English audio files, the corresponding transcribed text, and the Korean translation of the text. Audio files are provided for each TED talk, and the transcriptions and translations are provided for each subset with time information. Our corpus has the same structure as MuST-C [10]. We generated our training data automatically using the proposed corpus creation method from 3138 TED Talks, whereas the development and evaluation data were created manually. More detailed statistics are presented in Table 5.

5.2 | Baseline experiments

In this section, we present baseline results of the CAS-ST and E2E-ST models trained with EnKoST-C. All experiments were performed using the ESPnet toolkit [18] and Transformer models. Here, we briefly describe the architecture of each model. More detailed information is provided in the ESPnet-ST recipe for MuST-C [12,19].

Moreover, the E2E-ST model comprises a speech encoder and translation decoder. For speech features, we used 80-dimensional log-Mel filter-bank coefficients with three-dimensional pitch features, resulting in

TABLE 4 Examples of bilingual sentence alignment errors

Error types	Examples
1. Substitution	English: <i>Can you see?</i> Korean: 예, 좋습니다. [Good.]
2. Deletion	English: <i>This is true. I believe this.</i> Korean: 저는 믿습니다. [I believe this.]
3. Insertion	English: <i>I don't want to play anymore!</i> Korean: 그만! 더 안 해 봐도 돼! [No! I don't want to play anymore!]
4. Other	English: <i>Thank you.</i> Korean: (박수) [(Applause)]

Notes: Bracket indicate examples translated into English. Italics indicate misaligned sentences.

83-dimensional speech features. We used the 8 k vocabulary, comprising English and Korean vocabularies based on byte pair encoding units [33]. The speech encoder used 12 self-attention blocks stacked on two VGG blocks, and the translation decoder used six self-attention blocks.

In the CAS-ST system, the ASR model consists of a speech encoder and transcription decoder, and the MT model comprises a source text encoder and translation decoder, both of which used six self-attention blocks. Additionally, the source text of the encoder and transcription of the decoder used lowercase English transcription without punctuation marks.

The ASR performance was measured by the word error rate (WER) computed on lowercase, tokenized text without punctuation marks. The E2E-ST, CAS-ST, and MT results were computed using 4-g BLEU scores [34, 35]. Further, the BLEU score was measured with a single reference to Korean morpheme units tokenized using the Moses MT system [36].

Table 6 shows the WER and BLEU scores for ASR, MT, CAS-ST, and E2E-ST models. The CAS-ST model used ASR outputs with WERs of 9.9% and 7.4% as MT inputs. Consequently, the CAS-ST model underperformed the MT model, which uses English transcription as inputs. Here, all models were trained using only EnKoST-C.

TABLE 5 Statistic for each EnKoST-C subset

Subsets	Talks	Hours	Sentence pairs	Source tokens	Target tokens
train	3138	558.8	340K	6.0M	4.2M
dev	11	2.5	1585	26.5K	18.6K
tst-COMMON	27	4.0	2532	44.4K	29.9K
tst-HE	11	1.1	544	10.0K	7.0K

Note: Target tokens are measured in word–phrase units provided by Korean subtitles.

TABLE 6 Baseline results for ASR, MT, and ST models

Models		tst-COMMON	tst-HE
ASR (↓)		9.8	8.2
MT (↑)		17.1	18.9
CAS-ST (↑)		16.1	17.9
E2E-ST (↑)	Transformer	13.8	15.5
	+Speed perturbs	14.1	16.0
	+ASR init.	15.7	17.0
	+ASR/MT init.	15.5	17.3

Notes: The ASR results indicate word error rate where lower values are better. The other results indicate BLEU scores where higher values are better.

Furthermore, we present the baseline performance of the E2E-ST model trained with speed perturbation and a pretrained speech encoder and translation decoder. In Table 6, the “Transformer” is the vanilla E2E-ST model, which is directly trained on English speech data paired with Korean text translations. Here, the vanilla E2E-ST model showed improved performance when we used threefold speed perturbation by changing the speed with factors of 0.9, 1.0, and 1.1. Moreover, the highest E2E-ST BLEU scores were achieved when the model weights were initialized using the speech encoder of the ASR model and the translation decoder of the MT model.

The final E2E-ST model underperformed the CAS-ST model in terms of BLEU scores; however, the performance gap was reduced by the advanced modeling. Additionally, we expect that the E2E-ST model would outperform the CAS-ST model if larger amounts of ST data are collected using the proposed method of data collection.

6 | CONCLUSIONS

This paper introduced EnKoST-C, an EnKoST-C. This corpus contains 339 710 English–Korean sentence pairs aligned to 558.8 h of English speech. We also presented a detailed explanation of an automatic data collection method with an innovative sentence alignment method

that uses subtitle time information and the similarity measure based on bilingual sentence embedding. The proposed approach demonstrated the best sentence alignment performance with an F_1 score of 0.96. We also presented the baseline performance of an English–Korean E2E-ST model trained with EnKoST-C.

Meanwhile, we are releasing our corpus on an open data hub site of the Korean government, under a CC BY-NC-ND 4.0 International. To the best of our knowledge, EnKoST-C is the first publicly available EnKoST-C. We expect that this corpus will be widely used as a benchmark corpus for English–Korean ST.

ACKNOWLEDGMENTS

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

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest.

ORCID

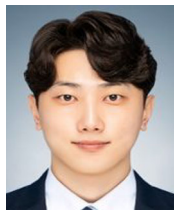
Jeong-Uk Bang  <https://orcid.org/0000-0002-0439-6802>

REFERENCES

1. A. Béard, O. Pietquin, L. Besacier, and C. Servan, *Listen and translate: A proof of concept for end-to-end speech-to-text translation*, (Proc. NIPS Workshop on end-to-end learning for speech and audio processing, Barcelona, Spain), Dec. 2016.
2. H. Ney, *Speech translation: Coupling of recognition and translation*, (Speech translation: Coupling of recognition and translation, Phoenix, AZ, USA), 1999, pp. 517–520.
3. H. Inaguma, K. Duh, T. Kawahara, and S. Watanabe, *Multilingual end-to-end speech translation*, (Multilingual end-to-end speech translation, Singapore), 2019, pp. 570–577.
4. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, *Librispeech: an asr corpus based on public domain audio books*, (Librispeech: an asr corpus based on public domain audio books, South Brisbane, Australia), Apr. 2015, pp. 5206–5210.
5. F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Estève, *TED-LIUM 3: Twice as much data and corpus reparation for experiments on speaker adaptation*, (Proc. International Conference on Speech and Computer, Leipzig, Germany), 2018, pp. 198–208.

6. J.-U. Bang, S. Yun, S. H. Kim, M. Y. Choi, M. K. Lee, Y. J. Kim, D. H. Kim, J. Park, Y. J. Lee, and S. H. Kim, *KsponSpeech: Korean spontaneous speech corpus for automatic speech recognition*, *Appl. Sci.* **10** (2020), no. 19, 6936.
7. A. C. Kocabiyikoglu, L. Besacier, and O. Kraif, *Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation*, (Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan), 2018.
8. B. Beilharz, X. Sun, S. Karimova, and S. Riezler, *LibriVoxDeEn: A corpus for German-to-English speech translation and German speech recognition*, (Proceeding of the 12th Language Resources and Evaluation Conference, Marseille, France), 2020, pp. 3590–3594.
9. 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*, (Proceedings of the 10th International Workshop on Spoken Language Translation: papers, Heidelberg, Germany), 2013.
10. R. Cattoni, M. A. Di Gangi, L. Bentivogli, M. Negri, and M. Turchi, *MuST-C: A multilingual corpus for end-to-end speech translation*, *Comput Speech Lang* **66** (2021), 101155.
11. R. S. Wurman and H. Marks, *TED talks*, 1984, <https://www.ted.com/talks/> [last accessed September 2020].
12. H. Inaguma, S. Kiyono, K. Duh, S. Karita, N. E. Soplín, T. Hayashi, and S. Watanabe, *ESPnet-ST: All-in-one speech translation toolkit*, (Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations), 2020, pp. 302–311.
13. C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, and J. Pino, *Fairseq S2T: Fast speech-to-text modeling with Fairseq*, (Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations, Suzhou, China), 2020, pp. 33–39.
14. L. C. C. Rosado, *Cinema at the service of natural language processing*, M.S. thesis, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal, 2016.
15. D. Varga, P. Halácsy, A. Kornai, V. Nagy, L. Németh, and V. Trón, *Parallel corpora for medium density languages*, In *Amsterdam studies in the theory and history of linguistic science*, Series 4, Benjamins, Amsterdam, 2007.
16. F. Braune and A. Fraser, *Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora*, (Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Beijing, China), 2010, pp. 81–89.
17. B. Thompson and P. Koehn, *Vecalign: Improved sentence alignment in linear time and space*, (Proc. Conference Empirical Methods Natural Language Processing-International Joint Conference on Natural Language Processing, Hong Kong, China), 2019, pp. 1342–1348.
18. S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N.E. Yalta Soplín, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, *Espnet: End-to-end speech processing toolkit*, arXiv preprint, 2018. <https://doi.org/10.48550/arXiv.1804.00015>
19. H. Inaguma, N. Kamo, S. Watanabe, and Y. Hayashibe, *ESPnet MuST-C recipe*, https://github.com/espnet/espnet/tree/master/egs/must_c/st1/2020 [last accessed September 2020].
20. W. I. Cho, S. M. Kim, H. Cho, and N. S. Kim, *Kosp2e: Korean speech to English translation corpus*, (Proc. Interspeech, Brno, Czechia), 2021, pp. 3705–3709.
21. R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, *How2: A large-scale dataset for multimodal language understanding*, (Proc. Conference on Neural Information Processing Systems, Montreal, Canada), 2018.
22. C. Federmann and W. D. Lewis, *Microsoft speech language translation (MSLT) corpus: The iwslt 2016 release for English, French and German*, (Proc. International Conference on Spoken Language Translation, Seattle, WA, USA), 2016.
23. C. Wang, A. Wu, and J. Pino, *Covost 2: A massively multilingual speech-to-text translation corpus*, arXiv preprint, 2020. <https://doi.org/10.48550/arXiv.2007.10310>
24. M. Brayan, *Figure Eight website*, 1996, <https://appen.com/> [last accessed September 2020].
25. O. F. Zaidan and C. Callison-Burch, *Crowdsourcing translation: Professional quality from non-professionals*, (Proc. Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA), 2011.
26. R. M. Ochshorn and M. Hawkins, *Gentle*, 2017, <https://github.com/lowerquality/gentle/> [last accessed September 2020].
27. W. A. Gale and K. Church, *A program for aligning sentences in bilingual corpora*, *Comput Linguist* **19** (1993), no. 1, 75–102.
28. J. Tiedemann, *Improved sentence alignment for movie subtitles*, (Proc. Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria), 2007, pp. 582–588.
29. F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, *Language-agnostic bert sentence embedding*, arXiv preprint, 2020. <https://doi.org/10.48550/arXiv.2007.01852>
30. B. Fabrice, and M. Niedermayer, *Ffmpeg*, 2012, <https://www.ffmpeg.org/> [last accessed September 2020].
31. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, *The Kaldi speech recognition toolkit*, (Proc. ASRU), 2011.
32. P. Bell, M. J. F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, A. Renals, O. Saz, M. Wester, and P. C. Woodland, *The MGB challenge: Evaluating multi-genre broadcast media recognition*, (Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, Scottsdale, AZ, USA), 2015, pp. 687–693.
33. T. Kudo, *Subword regularization: Improving neural network translation models with multiple subword candidates*, (Proc. 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia), 2018, pp. 66–75.
34. K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, *Bleu: A method for automatic evaluation of machine translation*, (Proc. 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA), 2002, pp. 311–318.
35. R. Sennrich and J. Barry, *BLEU score*, 2017, <https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu-detok.perl/> [last accessed September 2020].
36. E. L. Park and S. Cho, *KoNLPy: Korean natural language processing in Python*, (Proc. Annual Conference on Human and Language Technology), 2014, pp. 133–136.

AUTHOR BIOGRAPHIES



Jeong-Uk Bang received a BS degree in electronics engineering, an MS degree in control and instrumentation engineering, and a PhD degree in control and robot engineering at Chungbuk National University, Cheongju, Rep. of Korea, in 2013, 2015, and 2020, respectively. He is currently a post-doctoral researcher at the Artificial Intelligence Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea. His research interests are speech recognition, speech translation, end-to-end model, and speech data refinement.



Joon-Gyu Maeng received a BS degree in computer engineering from Hanbat National University, Daejeon, Rep. of Korea, in 2020. He is currently studying for an MS degree at the University of Science and Technology, Daejeon, Rep. of Korea, and is a student researcher at the Artificial Intelligence Research Laboratory, Electronics, and Telecommunications Research Institute, Daejeon, Rep. of Korea. His research interests are speech recognition, speech detection, speech translation, and end-to-end models.



Jun Park received BS and MS degrees in electronics engineering at Seoul National University in 1981 and 1983, respectively, and a PhD degree in electrical engineering systems at the University of Southern California in 1994. He is currently a special fellow at the Artificial Intelligence Research Laboratory, Electronics, and Telecommunications Research Institute, Daejeon, Rep. of Korea. His

research interests are speech recognition and speech translation.



Seung Yun received a PhD degree in computer software from the University of Science and Technology, Daejeon, Rep. of Korea. He is currently a principal researcher with the Artificial Intelligence Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea. His current research interests include the design of speech databases, language understanding, human-machine interface, speech recognition, and speech translation.



Sang-Hun Kim received a BS degree in electrical engineering from Yonsei University, Seoul, Rep. of Korea, in 1990, an MS degree in electrical engineering and electronic engineering from KAIST, Daejeon, Rep. of Korea, in 1992 and a PhD degree in electrical, electronic, and information communication engineering from the University of Tokyo, Japan, in 2003. Since 1992, he has been working for the Electronics and Telecommunications Research Institute, Daejeon, South Korea. His research interests are speech translation, spoken language understanding, and multimodal information processing.

How to cite this article: J.-U. Bang, J.-G. Maeng, J. Park, S. Yun, and S.-H. Kim, *English-Korean speech translation corpus (EnKoST-C): Construction procedure and evaluation results*, ETRI Journal (2022), 1–10. <https://doi.org/10.4218/etrij.2021-0336>