




Spoken-to-written text conversion for enhancement of Korean–English readability and machine translation

HyunJung Choi²  | Muyeol Choi¹ | Seonhui Kim² | Yohan Lim²  |
Minkyu Lee¹ | Seung Yun¹ | Donghyun Kim¹  | Sang Hun Kim¹

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

²Department of Artificial Intelligence,
University of Science and Technology,
Daejeon, Republic of Korea

Correspondence

Sang Hun Kim, Integrated Intelligence
Research Section, Electronics and
Telecommunications Research Institute,
Daejeon, Republic of Korea.
Email: ksh@etri.re.kr

Funding information

Electronics and Telecommunications
Research Institute, Grant/Award Number:
23ZS1100

Abstract

The Korean language has written (formal) and spoken (phonetic) forms that differ in their application, which can lead to confusion, especially when dealing with numbers and embedded Western words and phrases. This fact makes it difficult to automate Korean speech recognition models due to the need for a complete transcription training dataset. Because such datasets are frequently constructed using broadcast audio and their accompanying transcriptions, they do not follow a discrete rule-based matching pattern. Furthermore, these mismatches are exacerbated over time due to changing tacit policies. To mitigate this problem, we introduce a data-driven Korean spoken-to-written transcription conversion technique that enhances the automatic conversion of numbers and Western phrases to improve automatic translation model performance.

KEYWORDS

automatic machine translation, speech recognition, spoken-to-written conversion

1 | INTRODUCTION

Owing to the incredible advancements in neural networks models that support end-to-end automatic speech recognition (ASR), global users now expect high-quality, instantaneous speech transcriptions that support enhanced readability. Conventional speech-to-text training corpora for ASR models are applied text normalization procedures. However, because the transcription corpora for end-to-end ASR methods employs written forms, text normalization is largely unnecessary. This technical shift has led to the mixed presence of spoken and written language outputs within the transcription corpora, which results in inconsistent outcomes. Korean characters are applied as separate written (formal) and spoken (phonetic) forms. The written form represents the orthography of Korean words and is

employed in both reading and writing, particularly in official documents.

In contrast, the spoken form represents the physical pronunciation of words and phrases and can be visualized as sound representations using consonants and vowels. Notably, as with most languages, the spoken form can differ drastically from the written form, leading to ambiguities, particularly regarding numbers and Western words and phrases embedded in the Korean language. This ambiguity is a major source of errors in automatic translation models. For example, consider the phrase “100번 문제 보여주세요.” The expression “100번” in the written form and “백 번” in the spoken form are both valid. However, if the sentence “백 번 문제 보여주세요” is provided to an automatic translator, it will be translated incorrectly as “show me

the problem one hundred times,” instead of the intended “show me problem 100.” Additionally, the term “IBM” is transcribed in Korean as “아이비엠.” However, when the sentence “아이비엠 건물로 와주세요” is input into an automatic translator, it often produces the inaccurate translation, “come to the Ivy M building.” This stems from the inherent difficulty of machine-learning models in distinguishing potential double meanings when numbers and Western terms are transcribed in the Korean spoken form. Moreover, when using automatic Korean–English translation tools with Korean ASR outputs, these errors occur frequently.

These problems can arise from the construction of an ASR transcription training corpus. Noting that audio from broadcasts, lectures, news, and dramas are commonly used to create these datasets, their transcriptions into captions for the hearing-impaired have been used to pair with the written forms [1, 2]. This process saves time and costs compared with new manual transcriptions. However, the creation of captions for hearing-impaired persons primarily mimic the written form, which can conflict with the original spoken form. Moreover, numbers and Western words and phrases exhibit very low occurrence frequencies due to their simplification and variation. Although these combinations do not affect human semantic speech recognition, they can undermine automatic translation models based on the various ambiguities. Therefore, there is a strong demand for a model that ensures a bountiful and resilient ASR transcription corpus.

Text normalization involves converting written text into spoken form. Within this paradigm, number normalization primarily relies on a rule-based method as their reading variations depend on context. Recent advances have led to a transition from rule-based to end-to-end modeling using neural networks. Research into Korean number normalization has included transformers [3] and end-to-end text normalizers to enhance Korean speech synthesis [4]. Inverse text normalization (ITN) is more

applicable to Western languages and is often applied to their ASR post-processing stages.

As such, ITN research has been predominantly conducted in English and other Western multilingual combinations. A hybrid method [5] that combines a neural network and a rule-based finite-state transducer [6] was developed to recognize and post-process digit patterns in speech recognition. Applying an ITN neural network involves numerical data augmentation [7, 8] to achieve multilingual ITN. As such, streaming transformer taggers [9] are commonly used to label vocabulary tokens, which enables stable conversions using weighted finite-state transducers [6]. However, tagging vocabulary tokens can lead to errors when the aforementioned contextual ambiguities are strong. A similar study involved speech-to-spoken and written text (S2SWT), focusing on generating parallel spoken and written forms through speech recognition [10, 11]. In addition to the ITN type, ASR post-processing has been categorized into the correction (COR) type. Recent papers on this topic are summarized in Table 1. However, as indicated, limited research is available on ITN concept for the Korean language, which involves conversion from spoken to written form.

To address these challenges, we introduce a Korean data-driven spoken-to-written (K-STW) transcription conversion model that automatically converts spoken forms into transcripts and to standardize the existing corpora into written form, which is the baseline format used for end-to-end ASR. Unlike speech-dependent ASR, this approach leverages the advantages of text corpora and offers the unique benefits of highly consistent transcriptions and enhances automatic translation performance. The architecture of the K-STW model is described in detail in Section 2, and Section 3 describes the process of constructing the training set for the model. Sections 4 and 5 provide analyses of the experimental results and performance, respectively, and conclusions are drawn in Section 6.

TABLE 1 ASR post-processing categorization.

Category	Model		Database		Language	Institute
	ASR	LM	Speech	Text		
COR	LAS	LSTM	LibriSpeech	Non	English	Google [12]
	Transformer	N-gram/TXL				NVIDIA [13]
	Transformer		TED/AIHub		Korean	Korea Univ [14]
ITN	Transformer + BERT		Non	Wikipedia/News-C/MuST-C	English, German, Spanish, Italian	Amazon [5]
	S2S bi-LSTM			Social Media Corpus	English, French, Italy, Spanish	META [7,8]
	Transformer	Tagger+WFST	Non-public data		English	Microsoft [9]
	Transformer		Corpus of Spontaneous Japanese(CSJ)		Japanese	NTT [10]

2 | PROPOSED K-STW MODEL

Over the past 30 years, the Electronics and Telecommunications Research Institute (ETRI) has substantially contributed to the development of Korean speech datasets. The ETRI Manual Transcription Rule [15], which is their dataset construction standard, was designed to incorporate a wide array of environmental variables into speech data using diverse tags. Notably, Korean pronunciation in both spoken and written forms is recorded in a dual-transcription format. Additional efforts have been made to accurately transcribe numbers and Western phrases by considering their various contexts. For instance, in the written form of “6 am,” numeral six indicates them time, which should be pronounced as “여섯” in the ordinal form, rather than “육” in the cardinal form. To address this particular issue, the accurate pronunciation “여섯” is indicated after “6,” as in “오전 (6)(여섯)시면 어김없이 일어나 밥을 짓는다.” Similar dual transcriptions are applied to Western text, as in

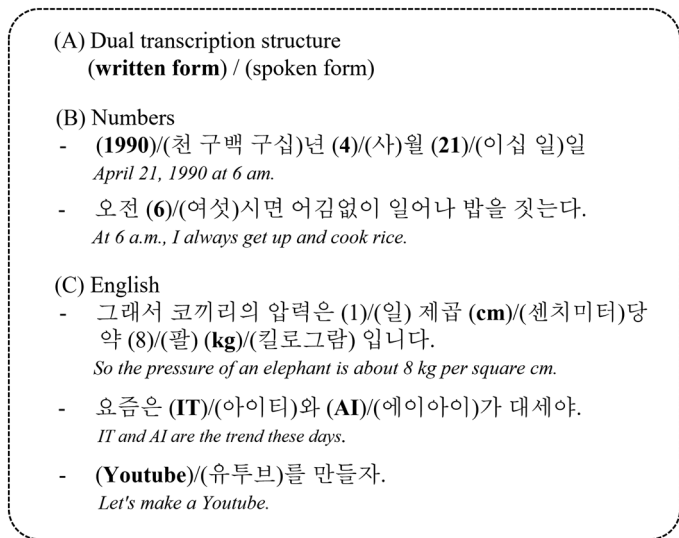


FIGURE 1 Examples of dual transcription: (A) dual structure, (B) numbers, and (C) Western text.

“요즘은 (IT)/(아이티)와 (AI)/(에이아이)가 대세야.” More examples are shown in Figure 1.

This dual-transcription dataset enables a straightforward distinction between numbers, Western text, and other text types. Notably, its structure resembles that of a machine translation model in which the parallel corpus structure of spoken and written forms facilitates training. Hence, our data-driven ITN K-STW method integrates a transformer [16] with an encoder–decoder architecture to transform input spoken-form tokens into output written-form varieties. This conversion occurs through a series of operations facilitated by the transformer. The structural synergy of the model enables the smooth conversion of spoken forms into written text, which is expected to support transcription and translation tasks.

Figure 2 illustrates the K-STW model process, showing how spoken-form tokens are processed as inputs and transformed by the decoder into their corresponding written-form tokens as outputs. For example, when input tokens “열두 시 십 분” are fed to the encoder, the model separately processes them. Among these tokens, those representing numbers, “열두” and “십,” are meticulously converted into their written forms, “12” and “10,” respectively. Hence, they subsequently emerge as the decoder’s output. This intricate transformation reflects the model’s seamless translation capability. The training process of the K-STW model leverages the use ESPnet machine translation (MT) script [17], which facilitates parameter and model refinements for performance optimization.

3 | TEXT REFINEMENT FOR K-STW MODEL TRAINING

A total of 8.6 M sentences were used to train the K-STW model, including 4.6 M from the AIHub Korean Lecture (AIHub-KL) speech dataset [18] and 4 M from the ETRI Korean Common (ETRI-KC) speech dataset [19]. The

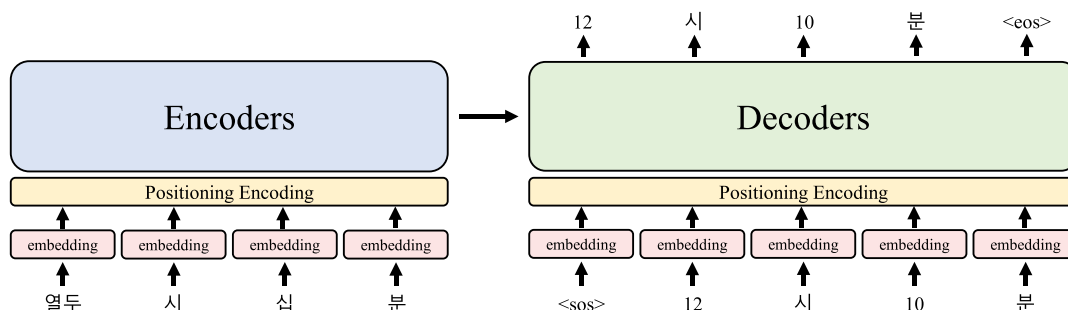


FIGURE 2 Diagram of the proposed K-STW model.

AIHub-KL dataset is a publicly and freely available repository, whereas the ETRI-KC dataset includes purchasing fees. Both datasets adopt a dual-transcription format encompassing both written and spoken forms of numbers and Western (primarily English) text.

The AIHub-KL dataset comprises audio extracted from broadcasts and online lectures from the Korean Educational Broadcasting System. These educational broadcasts cater to students and the public and are intended for learning purposes. Human non-experts listened to the content and transcribed it manually. Given the diverse array of numbers and English text in the content, this dataset is deemed to be strongly applicable for K-STW model training. However, transcription errors occurred due to inconsistent transcription rules applied during the manual transcription, performed by multiple individuals. Therefore, text refinement was necessary to mitigate these discrepancies. Detailed information on the AIHub-KL dataset is presented in Table 2.

The ETRI-KC dataset provides standardized written-form tokens that correspond to spoken forms by adhering to the conventions for the composition of newspaper articles. This dataset covers a diverse range of news article topics, including economics, society, science, and culture. Furthermore, it ensures that the content remains unbiased regardless of field, rendering it highly effective for model training. The goal of the proposed K-STW model is

to generate written forms from input spoken forms. However, with end-to-end ASR, output instances may include a written form as a recognition result. Therefore, assuming only the spoken form as the input is unrealistic. Instead, during training, we added inputs in written form, for which the output was designed for retention. Hence, the numbers and English text are handled correctly by the K-STW model.

Before training with the AIHub-KL and ETRI-KC datasets, refinements were performed to ensure that the model produced consistent results. This process involved addressing two aspects of dual transcription: number and English notations, as detailed in Sections 3.1 and 3.2.

3.1 | Number notations

Depending on the context, Arabic numerals “1, 2, 3, ...” can be represented in Sino-Korean words as “일, 이, 삼, ...” in their cardinal form, or in Korean native words as “하나, 둘, 셋, ...” in their ordinal form. Transcription by individuals who do not pay close attention to these different applications often result in errors [20]. For example, consider the following dual-transcription sentence: “서술형 (1)/(일) 번 문제 (1)/(한)번 풀어볼까요?” This sentence can be translated into English as “shall we solve narrative problem number 1?” If “1” is transcribed using both native Korean and Sino-Korean terms, semantic errors are introduced, leading to duplication errors. Initially, a semantic error arises when the written transcription is read as “서술형 문제 1번 문제 1번 풀어볼까요?” This sentence is translated into English as “shall we solve narrative problem number 1 problem number 1?” This translation duplicates the first problem, described above. Hence, the written form should be accurately adjusted from the spoken form, and during K-STW training, the spoken-form input tokens, “일” and “한,” are mapped to “1.” To prevent this conflict, the written forms of the native Korean words in the dual-transcription data must be modified by revising the dual transcription to “서술형 (1)/(일) 번 문제 (한)/(한) 번 풀어볼까요?” Depending on the intended numerical meaning, adjustments are made to the written form to correspond to the intended spoken meaning.

For Sino-Korean numerals, such as “일, 이, and 삼,” we refine the training data to ensure that the written form provides the correct numerals, “1, 2, and 3,” respectively. Furthermore, for native Korean numbers, we retain the original native notations for numbers up to 10 in their written form. For numbers 11 and above, the written form is changed to Arabic numerals to enhance readability. A threshold of 10 was therefore selected for convenience and clarity of communication. The criteria

TABLE 2 Details of AIHub-KL speech dataset.

Classification	Time (h)	Description of selected data
Elementary school	960+	5 subjects: Korean language, mathematics, social studies, science, history 640 h of Korean, 680 h of mathematics, 570 h of social studies, 610 h of science, 550 h of history
Middle school	750+	
High school	1340	
Vocational education/certificate	530+	10 categories: Korean language, Korean history, social studies, science, mathematics, professional qualifications, finance, management, information technology, technology 50 h–80 h per category
Others	420+	10 categories: humanities, philosophy, literature, art, science, social studies, information technology, education, etc. 30–100 h per category

for the numerical refinement in the training dataset are listed in Table 3.

3.2 | English notation

Owing to the content of online lectures, the corresponding corpus frequently includes terms that indicate units, such as weight, length, size, etc. Notably, these are often presented in abbreviated form. In the corpus, unit expressions are presented in written English, encompassing both their abbreviations and spelled-out forms. However, the spoken Korean equivalents of these expressions are manifest in diverse ways. We list the written and spoken forms of several units in Table 4 and categorize them according to their abbreviations, spelled-out English forms, and various Korean forms.

The Korean spoken equivalents of English-written forms exhibit variations in unit expression. For instance,

TABLE 3 Text refinement criteria for numbers.

	Spoken form	Written form
Sino-Korean numbers	일, 이, 삼, 사, 오 ... (1, 2, 3, 4, 5, ...)	1, 2, 3, 4, 5, ...
Native numbers (up to 10)	하나(한), 둘(두), 셋(세), 넷(네), 다섯, 여섯, 일곱, 여덟, 아홉, 열 (one, two, three, four, five, six, seven, eight, nine, ten)	Same as spoken form
Native numbers (11 and above)	열 하나(한), 열 둘(두), 열 셋(세), ... (eleven, twelve, thirteen, ...)	11, 12, 13, ...

TABLE 4 Classifications of expressions of units by their abbreviations, spelled-out English forms, and Korean pronunciations.

Written form	Spoken form
Abbreviation	Spelled-out English
kg	kilogram
cm	centimeter

TABLE 5 Classification of English words by selected English term, common variations, and Korean spoken forms.

Written form	Spoken form
Selected English	Used English
Wi-Fi	Wi-fi, wi-fi, WI-FI
YouTube	Youtube, youtube, YOUTUBE
UNESCO	Unesco, unesco
K-pop	K-Pop, k-pop, K-POP

“kilogram” has six distinct spoken Korean variations. Considering Korean spacing rules, further diversity is likely.

In the ETRI-KC dataset, units are consistently represented in standardized written English forms, predominantly using abbreviations. By contrast, the AIHub-KL dataset contains a mixture of abbreviations and spelled-out English forms. Hence, we adjusted the written English forms to uniformly present the units as abbreviated terms. For units not listed in Table 4, we applied a data-driven approach to maintain consistency.

Despite the existence of various spoken forms of Korean units, their correlation to written English forms was established using a dual-transcription corpus. Hence, multiple Korean-spoken forms can be converted into uniform English abbreviations, which illustrates the advantages of using the K-STW model. Uniform notations offer more stable and consistent conversion results than rule-based approaches, which require addressing every possible variation. The K-STW model is based on a dual-transcription corpus; hence, it offers notable benefits for achieving many-to-one mappings.

Similar to English unit notations, variations were observed for some English terms appearing in the corpus, as listed in Table 5. Unlike unit notations, the Korean spoken form is consistent, whereas considerable variations appear in the English-written form. In this case, the many-to-one mapping maintained by the K-STW model must be modified into a one-to-many mapping. This scenario leads to overfitting during model training. Thus, to mitigate this problem, variations in the written English forms must be reduced to improve readability and resolve possible translation errors related to proper nouns. Similar to how we handled the English unit notations, we adopted a data-driven approach to maintain consistency for English terms that are not

explicitly listed in Table 5. To select a consistent English word, the term with the highest frequency is preferred from among the alternatives available in the corpus.

4 | EXPERIMENT

Among the corpora used to train the K-STW model, an analysis of word proportions separated by spaces revealed that numbers and English text comprised 3.6 and 1.8% of the data, respectively. Similarly, owing to the shortage of numbers and English expressions within the test set, we selected samples to perform a comprehensive performance evaluation of the K-STW model. To construct this set, 1000 sentences not used for training were randomly chosen while ensuring that they contained numbers and English text. Thus, the test set consisted of 30% sentences containing numbers, 30% sentences containing English text, and 40% sentences containing Korean text only. A diagram of the K-STW model evaluation is shown in Figure 3.

The K-STW model was evaluated as follows:

- (1) Apply text refinement as described in Section 3 to the test set
- (2) Apply the spoken form from the refined text as input to the K-STW model and use the written form as the reference for model evaluation
- (3) Measure the K-STW model accuracy by comparing its results to the reference

The test set of 1000 sentences contained 15,674 words, including spaces. In addition, there were 1517 target tokens that the K-STW model converted from spoken to written forms. These tokens were categorized into numbers and English text, as listed in Table 6. The number of targets and their ratio per pattern showed that the numbers and English text tokens accounted for 63% and 37% of the converted samples, respectively.

The accuracy of the K-STW model was determined by calculating the ratio of the number of predicted target tokens to the total number of target tokens. The predicted

target tokens comprised a combined count of numbers and English language target tokens. The model accuracy results are listed in Table 7, where the K-STW model demonstrated an accuracy of 84.91% for number tokens and 76.47% for English word tokens, resulting in an average accuracy of 80.95%. The lower accuracy for English words was likely caused by the proportion of English samples within the training set being smaller than that of the numbers.

5 | TRANSLATION IMPROVEMENT USING K-STW MODEL

To evaluate the translation performance of the K-STW model, 100 sentences containing numbers and English text were selected from the test set. These sentences were translated, and we obtained bilingual evaluation under study (BLEU) [21] and BLEU with representations (BLUERT) scores [22], as listed in Table 8.

TABLE 6 Distribution of target and predicted tokens in the test set.

	Number	English	Total
No. target tokens	956	561	1517
No. predicted tokens	799	429	1228

TABLE 7 K-STW model accuracy.

	Numbers	English	Average
Accuracy (%)	84.91	76.47	80.95

TABLE 8 Translation performance comparisons of BLEU and BLEURT scores.

Translation source	BLEU	BLEURT
Spoken form	0.40	0.20
K-STW output	0.45	0.37
Written form	0.47	0.41

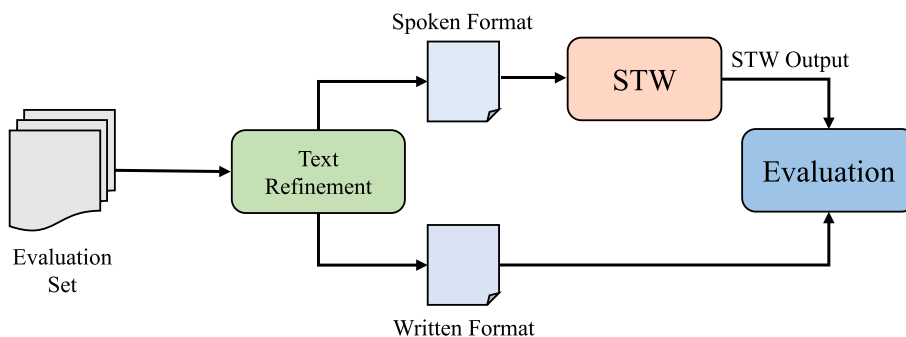


FIGURE 3 Diagram of the K-STW model evaluation.

TABLE 9 Examples of improved translation by applying the K-STW model.

		Translation result
Spoken form	제가 미국의 아이비엠 연구소에 갔습니다.	I went to the Ivy M Institute in the United States.
K-STW output	제가 미국의 IBM 연구소에 갔습니다.	I went to IBM Research in the United States.
Spoken form	비엠더블유 코리아는 천구백구십구년부터 해외 본사 인턴 프로그램을 운영하고 있다.	BMDoubleU Korea has been running an internship program for overseas headquarters since one thousand nine hundred and ninety-nine.
K-STW output	BMW 코리아는 1999년부터 해외 본사 인턴 프로그램을 운영하고 있다.	BMW Korea has been running an internship program at its overseas headquarters since 1999.
Spoken form	우리는 삼백 칠십 오 억 달러의 기름을 한 방울이라도 다 외국에서 사와야 하지 않으면 안되는 에너지 빈국의 나라	We're an energy-poor country that has to buy every single drop of our \$375 billion in oil from foreign countries.
K-STW output	우리는 375 억 달러의 기름을 한 방울이라도 다 외국에서 사와야 하지 않으면 안되는 에너지 빈국의 나라	We are an energy-poor country that has to buy every drop of our \$37.5 billion in oil from foreign countries.

For the performance comparison, we employed spoken form, the K-STW model output, and the written form for translation. For each translation source, we used the DeepL Korean–English translator [23]. To calculate the BLEU and BLEURT scores, we used a reference translation provided by a human expert and prepared three translation references containing numbers and English text in both spoken and written forms. Using the K-STW output, the BLEU score was 0.45, which is higher than the 0.40 for the spoken form and close to 0.47 for the written form. The BLEURT score was 0.37, which surpassed the performance of 0.2 for the spoken form. BLEURT allowed for more sophisticated quality evaluation, enabling a more detailed analysis of performance differences between models. Therefore, although the BLEURT score relatively decreased compared with the BLEU score, it can be concluded that the proposed model demonstrates better translation quality. The similarity in features between the BLEU and BLEURT scores suggests a certain degree of consistency between the two measurement methods. Hence, the K-STW model contributes to improved translation performance.

Table 9 presents three examples of enhanced translations supported by the K-STW model. Translations from the spoken form and the K-STW outputs were compared.

In the first example, the target tokens consisted of the written form “IBM” and spoken form “아이비엠.” During the translation of the spoken form, “아이비엠” was inaccurately translated to “Ivy M,” with the words “아이비” (IB) and “엠” (M), translated separately. When applying the K-STW model, the translation was accurate and retained the intended “IBM” representation.

The second example sentence mirrored the structure of the initial one, with the target token being “BMW” in its written form and “비엠더블유” in the spoken form. In addition, the numeric target token “1999” was present. From the translation of the spoken-form, “비엠더블유” was erroneously broken down into individual components “비” (B), “엠” (M), “더블” (double), and “유” (U), yielding “BMDoubleU.” In contrast, the application of the K-STW model ensured an accurate translation. The numeric target token “1999” was also translated correctly, even in the spoken form.

In the third example, target token “375” was presented in its written form, whereas the corresponding spoken form was “삼백칠십오” (three hundred seventy-five). This example illustrates the challenges arising from contextual influences on spoken forms. When solely applying the K-STW model, the translation accurately presented “\$37.5 billion” from the written form but erroneously generated “\$375 billion” from the spoken form.

These examples demonstrate that the K-STW model enhances translation performance and largely captures correct meanings.

6 | CONCLUSION

In this study, we proposed a data-driven K-STW model that automatically transforms spoken-form terms into their written forms. This model is intended to unify existing speech datasets into written forms to serve as a transcription format for end-to-end ASR. Notably, we addressed the automatic conversion performance for

numbers and Western (English) text and demonstrated a method that improves automatic translations. Our K-STW model provides an average conversion accuracy of 80.95%, as confirmed by comparative analyses and examples.

In any Korean text corpus containing Arabic numbers and English text, characters essential for written-form symbols, punctuation marks, and superfluous characters (interjections) often appear. These characteristics can either enhance or reduce readability, depending on the context. In future work, we plan to seek the enhancement of both readability and translation performance, including measures such as reinstating punctuation marks and omitting interjections.

ACKNOWLEDGMENTS

This study was supported by an Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean Government (23ZS1100, Core Technology Research for Self-improving Integrated Artificial Intelligence Systems).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ORCID

HyunJung Choi  <https://orcid.org/0009-0006-8947-6444>

Yohan Lim  <https://orcid.org/0009-0005-0007-4055>

Donghyun Kim  <https://orcid.org/0000-0002-2063-5551>

REFERENCES

- J.-U. Bang, M.-Y. Choi, S.-H. Kim, and O.-W. Kwon, *Automatic construction of a large-scale speech recognition database using multi-genre broadcast data with inaccurate subtitle time-stamps*, IEICE Trans. Inform. Syst. **103** (2020), no. 2, 406–415.
- 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 J. **45** (2023), no. 1, 18–27.
- J. Chun, C. Jo, J. Lee, and M.-W. Koo, *Number normalization in Korean using the transformer model*, KIISE **48** (2021), no. 5, 510–517.
- Y. Choi, Y. Jung, Y. Kim, Y. Suh, and H. Kim, *An end-to-end synthesis method for Korean text-to-speech systems*, Phonet. Speech Sci. **10** (2018), no. 1, 39–48.
- M. Sunkara, C. Shivade, S. Bodapati, and K. Kirchhoff, *Neural inverse text normalization*, (ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Canada), 2021, pp. 7573–7577.
- M. Mohri, *Weighted finite-state transducer algorithms. An overview*, Formal Lang. Appl. **2004** (2004), 551–563.
- L. Pandey, D. Paul, P. Chitkara, Y. Pang, X. Zhang, K. Schubert, M. Chou, S. Liu, and Y. Saraf, *Improving data driven inverse text normalization using data augmentation*, arXiv preprint, 2022, DOI [10.48550/arXiv.2207.09674](https://doi.org/10.48550/arXiv.2207.09674)
- D. Paul, Y. Pang, S.-J. Chen, and X. Zhang, *Improving data driven inverse text normalization using data augmentation and machine translation*, (Proc. Interspeech, Incheon, Rep. of Korea), 2022, pp. 5221–5222.
- Y. Gaur, N. Kibre, J. Xue, K. Shu, Y. Wang, I. Alphanso, J. Li, and Y. Gong, *Streaming, fast and accurate on-device inverse text normalization for automatic speech recognition*, (IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar), 2023, pp. 237–244.
- M. Ithori, H. Sato, T. Tanaka, R. Masumura, S. Mizuno, and N. Hojo, *Transcribing speech as spoken and written dual text using an autoregressive model*, (Proc. Interspeech, Dublin, Ireland), 2023, DOI [10.21437/Interspeech.2023-1655](https://doi.org/10.21437/Interspeech.2023-1655).
- M. Ithori, A. Takashima, and R. Masumura, *Parallel corpus for Japanese spoken-to-written style conversion*, (Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France), 2020, pp. 6346–6353.
- J. Guo, T. N. Sainath, and R. J. Weiss, *A spelling correction model for end-to-end speech recognition*, (ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK), 2019, pp. 5651–5655.
- O. Hrinchuk, M. Popova, and B. Ginsburg, *Correction of automatic speech recognition with transformer sequence-to-sequence model*, (ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain), 2020, pp. 7074–7078.
- C. Park, J. Seo, S. Lee, C. Lee, H. Moon, S. Eo, and H.-S. Lim, *BTS: back transcription for speech-to-text post-processor using text-to-speech-to-text*, (Proceedings of the 8th Workshop on Asian Translation (WAT2021)), 2021, pp. 106–116.
- 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, *Kspon-speech: Korean spontaneous speech corpus for automatic speech recognition*, Appl. Sci. **10** (2020), no. 19, DOI [10.3390/app10196936](https://doi.org/10.3390/app10196936).
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention is all you need*, Adv. Neural Inform. Process. Syst. **30** (2017).
- S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, and N. Chen, *Espnet: end-to-end speech processing toolkit*, arXiv preprint, 2018, DOI [10.48550/arXiv.1804.00015](https://doi.org/10.48550/arXiv.1804.00015)
- AIHub, *Aihub Korean lecture speech dataset*, 2020. Last accessed on August 27, 2023.
- ETRI, *Etri Korean common speech dataset*, 2004. Last accessed on August 27, 2023.
- Y.-I. Jung, J.-S. Kim, S.-H. Kim, Y.-J. Lee, and A.-S. Yoon, *A study on the arabic numeral reading rules in modern Korean*, (Annual Conference on Human and Language Technology. Human and Language Technology), 2002, pp. 16–23.
- M. Post, *A call for clarity in reporting bleu scores*, arXiv preprint, 2018, DOI [10.48550/arXiv.1804.08771](https://doi.org/10.48550/arXiv.1804.08771)
- T. Sellam, D. Das, and A. P. Parikh, *BLEURT: Learning robust metrics for text generation*, (Proceedings of Annual Meeting of the Association for Computational Linguistics), 2020. DOI [10.18653/v1/2020.acl-main.704](https://doi.org/10.18653/v1/2020.acl-main.704).
- D. Se, *DeepL translate: the world's most accurate translator*, 2017. <https://www.deepl.com/translator>

AUTHOR BIOGRAPHIES



HyunJung Choi received the BA degree in English language and literature (English for global communication) and media from Kookmin University, Seoul, Republic of Korea, in 2023. She is currently pursuing an M.S. degree at the University of Science and Technology, Daejeon, Republic of Korea, and is a student researcher at the Superintelligence Creative Research Laboratory at the Electronic and Telecommunication Research Institute, Daejeon, Republic of Korea. Her research interests include natural language processing, machine translation, and speech recognition.



Muyeol Choi received the BS degree in electronics and communication engineering from Dong-Eui University, Busan, Republic of Korea, in 1997, and the MS and PhD degrees in electronics engineering from Busan National University, Busan, Republic of Korea, in 1999 and 2011, respectively. From 2000 to 2003, he worked for Voiceware in Seoul, Republic of Korea. Since 2012, he has worked for the Superintelligence Creative Research Laboratory at the Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea. His primary research interests include speech recognition, speech translation, and spoken language understanding.



Seonhui Kim received the BS degree in computer software engineering from Dong-Eui University, Busan, Republic of Korea, 2022. She is currently pursuing an MS degree at the University of Science and Technology, Daejeon, Republic of Korea, and is a student researcher at the Superintelligence Creative Research Laboratory at the Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea. Her research interests include machine translation, natural language processing, speech recognition, and end-to-end modeling.



Yohan Lim received the BS degree in computer engineering from the School of Engineering, Chungnam National University, Daejeon, Republic of Korea, in 2019. He is currently pursuing an integrated MS and PhD degrees at the University of Science and Technology, Daejeon, Republic of Korea, and is a

student researcher at the Superintelligence Creative Research Laboratory at the Electronic and Telecommunication Research Institute, Daejeon, Republic of Korea. His research interests include multilingual speech recognition, speech synthesis, synthetic data augmentation, self-supervised learning, multimodal representations, and large language models.



Minkyu Lee received the BS degree in computer engineering from the School of Computer Engineering, Kyungpook National University, Daegu, Republic of Korea, in 2011, and the MS degree in engineering from the University of Science and Technology, Daejeon, Republic of Korea, in 2014. Since 2014, he has worked at the Electronics and Telecommunications Research Institute in Daejeon, Republic of Korea. His main research interests include end-to-end multimodal speech recognition and system-level integration of speech recognition systems.



Seung Yun received the PhD in computer software from the University of Science and Technology, Daejeon, Republic of Korea. He is currently a principal researcher at the Superintelligence Creative Research Laboratory at the ETRI in Daejeon, Republic of Korea. His research interests include artificial intelligence, speech recognition, natural language processing, speech databases, and speech translation.



Donghyun Kim received the BS and MS degrees in computer and communication engineering from Korea University, Seoul, Republic of Korea, in 1999 and 2004, respectively, and the PhD in computer science from Korea University, Seoul, Korea, in 2008. Since 2009, he has worked for the Superintelligence Creative Research Laboratory at the Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea. His main research interests include multilingual speech translation, multimodal spoken language processing, and self-supervised deep learning.



Sang Hun Kim received the BS degree in electrical engineering from Yonsei University, Seoul, Korea, in 1990, the MS degree in electrical engineering and electronic engineering from KAIST, Daejeon, Korea, in 1992, and the PhD degree from the Department of Electrical, Electronic and Information

Communication Engineering from the University of Tokyo, Japan, in 2003. Since 1992, he has worked for ETRI. His interests include speech translation, spoken language understanding, and multimodal information processing.

How to cite this article: H. Choi, M. Choi, S. Kim, Y. Lim, M. Lee, S. Yun, D. Kim, and S. H. Kim, *Spoken-to-written text conversion for enhancement of Korean-English readability and machine translation*, ETRI Journal **46** (2024), 127–136, DOI [10.4218/etrij.2023-0354](https://doi.org/10.4218/etrij.2023-0354)