DOI: 10.4218/etr2.12666

GUEST EDITORIAL

ETRI Journal WILEY

Special issue on speech and language AI technologies

Recent advancements in artificial intelligence (AI) have substantially improved applications that depend on human speech and language comprehension. Human speech, characterized by the articulation of thoughts and emotions through sounds, relies on language, a complex system that uses words and symbols for interpersonal communication. The rapid evolution of AI has amplified the demand for related solutions to swiftly and efficiently process extensive amounts of speech and language data. Speech and language technologies have emerged as major topics in AI research, improving the capacity of computers to comprehend text and spoken language by resembling human cognition. These technological breakthroughs have enabled computers to interpret human language, whether expressed in textual or spoken forms, unveiling the comprehensive meaning of the intentions, nuances, and emotional cues expressed by writers or speakers.

Electronics and Telecommunications Research Institute (ETRI) Journal is a peer-reviewed open-access journal launched in 1993 and published bimonthly by ETRI, Republic of Korea. It is intended to promote worldwide academic exchange of research on information, telecommunications, and electronics.

This special is devoted to all aspects and future research directions in the rapidly progressing subject of speech and language technologies. In particular, this special issue highlights recent outstanding results on the application of AI techniques to understand speech or natural language. We selected 12 outstanding papers on three topics of speech and language technologies. Below, we provide a summary of commitments to this special issue.

The first paper [1] "Towards a small language model powered chain-of-reasoning for open-domain question answering" by Roh and others focuses on open-domain question-answering tasks that involve a chain of reasoning primarily implemented using large language models. Emphasizing cost effectiveness, the authors introduce EffiChainQA, an architecture centered on the use of small language models. They employ a retrieval-based language model that is known to address the hallucination issue and incorporates up-to-date knowledge, thereby addressing common limitations of larger language models. In addition, they introduce a question decomposer that leverages a generative language model and is essential for enhanced chain of reasoning.

In the second paper in this special issue [2], "CR-M-SpanBERT: Multiple-embedding-based DNN Coreference Resolution Using Self-attention SpanBERT" by Jung, a model is proposed to incorporate multiple embeddings for coreference resolution based on the SpanBERT architecture. The experimental results show that multiple embeddings can improve the coreference resolution performance regardless of the employed baseline model, such as LSTM, BERT, and SpanBERT.

As automated essay scoring has evolved from handcrafted techniques to deep learning methods, holistic scoring has improved. However, assessing specific traits remains challenging because of the limited depth of existing methods to model dual assessments for holistic and multitrait tasks. To address this challenge, a paper in this special issue titled "Dual-Scale BERT using Multi-Trait Representations for Holistic and Trait-Specific Essay Grading" [3] by Cho and others explores comprehensive feedback while modeling the interconnections between holistic and trait representations. The authors introduce the DualBERT-Trans-CNN model, which combines transformer-based representations with a novel dual-scale BERT encoder at the document level. By explicitly leveraging multitrait representations in a multitask learning framework, they emphasize the interrelation between holistic and trait-based score predictions to improve accuracy.

The fourth paper in this special issue [4], "Named entity recognition using transfer learning and small human- and meta-pseudo-labeled datasets" by Bae and Lim, introduces a high-performance model for named entity recognition for written and spoken language. The authors use transfer learning to leverage the previously developed KorBERT model as the baseline to overcome the challenges related to labeled data scarcity and domain shifts. They also adopt a meta-pseudo-label method using a teacher/student framework with labeled and unlabeled data. Their model presents two innovations: the combination of loss functions from human- and pseudo-labeled data and the updating of the teacher model only when a threshold is not reached.

While deep learning approaches are of keen interest, combining and applying them to traditional language analysis is also worthy, especially to explain analysis outcomes. The fifth paper in this special issue [5], “Transformer-Based Reranking for Improving Korean Morphological Analysis Systems” by Ryu and others, introduces this approach to Korean morphological analysis by combining dictionary-based techniques with transformer-based deep learning models. In particular, they use the BERT-based reranking system that substantially enhances the accuracy of the traditional dictionary-based morphological analysis methods. Their results demonstrate considerable performance improvements and highlight advantages of combining analytical and probabilistic models for language processing applications.

The sixth paper in this special issue [6], “Framework for evaluating code generation ability of large language models” by Yeo and others, introduces a systematic framework for evaluating the code generation capabilities of large language models and presents the derivation of a new metric called *pass-rate@n*, which captures granular accuracy levels by considering test pass rates. The experimental results demonstrate the effectiveness of the evaluation framework, which can be integrated with real-world coding platforms.

Another notable contribution to this field is presented in the paper titled “KMSAV: Korean multi-speaker spontaneous audiovisual dataset” by Park and others [7]. This paper presents a rich and extensive database encompassing approximately 150 h of rigorously transcribed and annotated audio-visual data supplemented by a diverse trove of 2000 h of untranscribed YouTube videos. This open-access corpus, accompanied by a tailored open-source framework, is validated through an evaluation using cutting-edge automatic and audio-visual speech recognition techniques.

The application of speech and language AI techniques to the clinical and medical domains has gathered research interest. The eighth paper [8], “Alzheimer's disease recognition from spontaneous speech using large language models” by Bang and others, presents the innovation of using large language models for predicting Alzhemier's disease by extensively using evaluation feedback generated by ChatGPT from image descriptions provided by potential patients. The feedback is used as an additional feature for speech multimodal transformer blocks. Experimental results demonstrate substantial improvements by leveraging the evaluation feedback from ChatGPT, thereby motivating the use of large language models for diagnosing some diseases.

The ninth paper [9], “Joint streaming model for backchannel prediction and automatic speech recognition” by Choi and others, addresses a crucial aspect of human conversation: the timely use of conversation backchannels such as “uh-huh” or “yeah.” This paper introduces a novel method that combines backchannel prediction with real-time speech recognition using a streaming transformer and multitask learning. The results show substantial improvements over existing methods, particularly in streaming scenarios, marking a substantial advancement toward more natural and engaging human–machine interactions.

The use of high-quality and adequate data for addressed application tasks is key to achieve stable high performance. The tenth paper in this special issue [10], “Spoken-to-written text conversion for enhancement of Korean–English readability and machine translation” by Choi and others, addresses the problem that Korean text produced by automatic speech recognition is often not presented in the written but in the spoken form, particularly when including numeric expressions and English words. Consequently, frequent ambiguities occur in similar types of errors for automatic speech translation. To mitigate these common types of errors, the authors propose a Korean spoken-to-written transcription conversion method trained on a large-scale dataset containing 8.6 million sentences formatted in a transcription style that aligns the written and spoken forms of text segments. Using the transcription conversion, substantial improvements in automatic speech translation from Korean to English are achieved, demonstrating the importance of high-quality task-aware data for properly training AI models.

The landscape of multimodal speech recognition has been drastically reshaped by the latest breakthroughs in deep learning. The paper titled “Multimodal Audiovisual Speech Recognition Architecture Using a Three-feature Multifusion Method for Noise-robust Systems” by Jeon and others addresses challenges of speech recognition in diverse noisy environments [11]. This paper presents an audio-visual speech recognition model that emulates human dialogue recognition, showing remarkable robustness across synthesized environments at nine different noise levels. By integrating audio and visual elements through a dense spatial–temporal convolutional neural network, the model achieves a substantially lower error rate than traditional methods. This study may pave the way for enhanced speech recognition services with both stability and improved recognition rates in noisy environments.

Language tutoring systems for nonnative speakers have taken a significant leap forward with the development of advanced end-to-end methods for automatic speech recognition and proficiency evaluation, as presented in the paper [12], "AI-based language tutoring systems with end-to-end automatic speech recognition and proficiency evaluation" by Kang and others. This paper details the creation of systems that proficiently assess and provide feedback on pronunciation and fluency using a combination of semisupervised and transfer learning techniques with diverse speech data. Highlighting its practical application, this study showcases two deployed systems, EBS AI PengTalk and KSI Korean AI Tutor, which enhance language learning for Korean elementary students and foreigners learning Korean, respectively.

The guest editors would like to thank all the authors, reviewers, and editorial staff of ETRI Journal for making this special issue successful. We are pleased to have been a part of the effort to timely publish high-quality technical papers. The presented studies on speech and language models will certainly contribute to the design and implementation of future AI systems.

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest.

Dong-Jin Kim [1]
Hyung-Min Park [2]
Harksoo Kim [3]
Seung-Hoon Na [4]
Gerard Jounghyun Kim [5]

[1]*Department of Data Science, Hanyang University, Seoul, Republic of Korea*
[2]*Department of Electronic Engineering, Sogang University, Seoul, Republic of Korea*
[3]*Department of Computer Science and Engineering, Konkuk University, Seoul, Republic of Korea*
[4]*Department of Computer Science & Artificial Intelligence, Chonbuk National University, Jeonju, Republic of Korea*
[5]*Department of Computer Science and Engineering, Korea University, Seoul, Republic of Korea*

Correspondence
Dong-Jin Kim, Department of Data Science, Hanyang University, Seoul, Republic of Korea.
Email: djdkim@hanyang.ac.kr

REFERENCES

1. J. Roh, M. Kim, and K. Bae, *Towards a small language model powered chain-of-reasoning for open-domain question answering*, ETRI J. **46** (2024), no. 1, DOI 10.4218/etrij.2023-0355
2. J. Jung, *CR-M-SpanBERT: Multiple-embedding-based DNN coreference resolution using self-attention SpanBERT*, ETRI J. **46** (2024), no. 1, DOI 10.4218/etrij.2023-0308
3. M. Cho, J. X. Huang, and O.-W. Kwon, *Dual-scale BERT using multi-trait representations for holistic and trait-specific essay grading*, ETRI J. **46** (2024), no. 1, DOI 10.4218/etrij.2023-0324.
4. K. Bae and J.-H. Lim, *Named entity recognition using transfer learning and small human- and meta-pseudo-labeled datasets*, ETRI J. **46** (2024), no. 1, DOI 10.4218/etrij.2023-0321.
5. J. Ryu, S. Lim, O.-W. Kwon, and S.-H. Na, *Transformer-based reranking for improving Korean morphological analysis systems*, ETRI J. **46** (2024), no. 1, 10.4218/etrij.2023-0364.
6. S. Yeo, Y.-S. Ma, S. C. Kim, H. Jun, and T. Kim, *Framework for evaluating code generation ability of large language models*, ETRI J. **46** (2024), no. 1, DOI 10.4218/etrij.2023-0357.
7. K. Park, O. Changhan, and S. Dong, *KMSAV: Korean multi-speaker spontaneous audiovisual dataset*, ETRI J. **46** (2024), no. 1, DOI 10.4218/etrij.2023-0352.
8. J.-U. Bang, S.-H. Han, and B.-O. Kang, *Alzheimer's disease recognition from spontaneous speech using large language models*, ETRI J. **46** (2024), no. 1, DOI 10.4218/etrij.2023-0356.
9. Y.-S. Choi, J.-U. Bang, and S. H. Kim, *Joint streaming model for backchannel prediction and automatic speech recognition*, ETRI J. **46** (2024), no. 1, DOI 10.4218/etrij.2023-0358.
10. H. J. 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 J. **46** (2024), no. 1, DOI 10.4218/etrij.2023-0354.
11. S. Jeon, J. Lee, D. Yeo, Y.-J. Lee, and S. J. Kim, *Multimodal audiovisual speech recognition architecture using a three-feature multifusion method for noise-robust systems*, ETRI J. **46** (2024), no. 1, DOI 10.4218/etrij.2023-0266.
12. B. O. Kang, H.-B. Jeon, and Y. K. Lee, *AI-based language tutoring systems with end-to-end automatic speech recognition and proficiency evaluation*, ETRI J. **46** (2024), no. 1, DOI 10.4218/etrij.2023-0322.

AUTHOR BIOGRAPHIES

Dong-Jin Kim received his BS, MS, and PhD degrees in electrical engineering from KAIST, Daejeon, Republic of Korea, in 2015, 2017, and 2021, respectively. He was a postdoctoral researcher in electrical engineering and computer sciences at UC Berkeley in 2022. He is an assistant professor at Hanyang University. He was a research intern with the Visual Computing Group, Microsoft Research Asia (MSRA). He was awarded a silver prize from Samsung Humantech paper awards and Qualcomm Innovation awards. His research interests include data issues in computer vision, especially multimodal learning problems.

Hyung-Min Park received his BS, MS, and PhD degrees in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, in 1997, 1999, and 2003, respectively. From 2003 to early 2005, he was a postdoctoral researcher at the Department of Biosystems, KAIST. From 2005 to early 2007, he was with the Language Technologies Institute, Carnegie Mellon University. In 2007, he joined the Department of Electronic Engineering, Sogang University, Seoul, Republic of Korea, where he is currently a professor. His main research interests include robust speech recognition and computer vision.

Harksoo Kim is a full professor at the Department of Computer Science and Engineering, Konkuk University. He received his BA degree in computer science from Konkuk University in 1996, MS degree in computer science from Sogang University in 1998, and PhD degree in computer science with a major in natural language processing from Sogang University in 2003. He visited the Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, as a research fellow in 2004. In 2005, he worked for the Electronics and Telecommunications Research Institute (ETRI) as a senior researcher. From March 2006 to February 2020, he worked for Kangwon National University as a professor. His research interests include natural language processing, dialogue understanding, information retrieval, and question answering.

Seung-Hoon Na received his PhD degree in computer science from POSTECH in 2008. Currently, he is a professor at the Department of Computer Science & Artificial Intelligence, Jeonbuk National University. Previously, he was a senior researcher at ETRI, Republic of Korea, after working in the School of Computing, Natural University of Singapore. He serves as a standing reviewer of Computational Linguistics, chair of the Special Interest Group of Human and Cognitive Language Technology, Republic of Korea, and reviewer for top-tier conferences on artificial intelligence, such as ACL, NAACL, COLING, EMNLP, AAAI, ICLR, and Neurips. He also served as a publication co-chair at COLING 2022. His research interests include natural language processing, information retrieval, and machine learning.

Gerard J. Kim received his BS in electrical and computer engineering from Carnegie Mellon University in 1987 and MS and PhD degrees in computer science from the University of Southern California in 1994. He is a professor of computer science and engineering at Korea University. His main research interests include human–computer interaction, virtual/mixed reality, and computer music. He serves as a section editor of the High-Performance Computing area for ETRI Journal.