Lexical and Semantic Evaluation of Whisper-Based Automatic Transcription of Korean Child-Directed Speech

Ioana Buhnila[1], Jiho Lee[2], Sung Hyu Chon[2], Jae-Hun Jung[2], Eon-Suk Ko[1],

[1]Center for Data Science in Humanities (CDS), Chosun University
ioana.buhnila@chosun.ac.kr; eonsukko@chosun.ac.kr

[2]MINDS, Pohang University of Science and Technology (POSTECH)
jihlee@postech.ac.kr; shhchon@postech.ac.kr; jung153@postech.ac.kr

Automatic Speech Recognition (ASR) tools such as Whisper (Radford et al., 2023) and WhisperX (Bain et al., 2023) have shown good performance when used for the transcription of standard language. However, Whisper's performance drops significantly when dealing with different speech types and noisy recording conditions. Automatically transcribing child-directed speech (CDS) recorded in natural home settings is a challenging task, as the presence of multiple speakers, TV speech, and specific particularities of CDS (word repetition, singing-like speech and various onomatopoeia) can trigger *hallucinations* in Whisper transcriptions (erroneous speech recognition). Traditional metrics such as Word Error Rate (WER) and Character Error Rate (CER) are widely used for evaluating ASR tools' performance. Nevertheless, relying only on these standard metrics for Korean is not sufficient, as they were designed for the English language and they do not adapt well to the agglutinative morphology of the Korean language. Small inflectional differences can substantially inflate token-level error rates despite partial semantic overlap.

This study conducted a fine-grained analysis of the surface-level differences between human and WhisperX transcription conventions, such as spacing, punctuation, and writing conventions. We developed a pre-processing pipeline to automatically clean and normalize the transcripts, taking the human annotation as reference. Moreover, we identified two major categories of ASR errors for Korean caregiver-child interactions: (1) *transcription scope mismatch* (differences between the segmented human annotation and the automatic transcription, background noise, TV or media sounds, and onomatopoeia (such as infant vocalizations) can be transcribed differently or can be completely absent); (2) *ASR output errors* (genuine speech recognition failures attributable to the tool, such as *hallucinated* repetitions of infant babbling, phonetic or lexical substitution, boilerplate *hallucinations*). These transcription errors, together with the transcription normalization issues, contributed to high WER and CER between human and WhisperX transcriptions.

Furthermore, we examined whether lexical and embedding-based semantic similarity metrics provide complementary evidence for evaluating WhisperX transcription quality in Korean CDS. We compared the human transcripts of 60 audio files of 5-min child-centered conversations recorded in Korean home environments (McDonald et al., 2021) with the automatically generated transcripts by the WhisperX turbo model (809M parameters). We used several NLP metrics, such as BLEU, TF-IDF, BERTscore (Zhang et al., 2019), and Sentence-BERT (Ham et al., 2020; Reimers and Gurevych, 2019) to capture the lexical and semantic similarity between the two types of transcripts. We report results on data after the first level of pre-processing,

which involved normalizing repeated and *hallucinated* onomatopoeia and vocalizations. While lexical scores are very low (BLEU score average of 0.17 on the full dataset), and TF-IDF scores demonstrate that only 60% on average of top-20 keywords overlapped between human and WhisperX transcripts, the average overall Sentence-BERT (0.78), and BERTscore F1 scores (0.76) show meaningful semantic similarity. However, these semantic scores have to be further analyzed, considering that automatic transcriptions of Korean CDS, an informal speech register, can include *hallucinations* or lexical substitutions, on top of different transcription practices of the Korean language.

To conclude, this study's findings suggested that despite extremely low metrics based on surface-forms, embedding-based semantic similarity between human and WhisperX transcripts of Korean caregiver-child conversations indicates that both transcripts convey similar meaning. Nevertheless, the overall semantic similarity score has to be completed with a fine-grained analysis of sentence-to-sentence embedding-based semantic scores to get a clearer picture of the exact semantic and pragmatic match. Moreover, while Sentence-BERT or BERTScore capture meaning according to token embeddings, these metrics do not fully represent the pragmatic dimensions of child-directed speech. Further pragmatic and Korean culture-centered analysis could be conducted using Large Language Models (LLMs) and Chain-of-Though or reasoning prompting to determine whether the caregiver-child interaction dynamic is accurately represented in both transcriptions.

References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. Kornli and korsts: New benchmark datasets for Korean natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 422–430.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Margarethe McDonald, Taeahn Kwon, Hyunji Kim,Youngki Lee, and Eon-Suk Ko. 2021. Evaluating the language environment analysis system for Korean. *Journal of Speech, Language, and Hearing Research*, 64(3):792–808.774.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.