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 strong performance in standard-language transcription tasks. However, Whisper's performance can decline substantially when applied to non-standard speech registers and noisy recording conditions. Automatically transcribing child-centered recordings in natural home settings is particularly challenging, as the presence of multiple speakers, TV speech, and specific properties of child-directed speech (CDS) such as word repetition, word play, and various onomatopoeia can trigger hallucinations or other erroneous speech-recognition outputs in Whisper transcriptions. Traditional metrics for evaluating ASR performance, including Word Error Rate (WER) and Character Error Rate (CER), were developed primarily with English in mind and do not readily accommodate Korean's agglutinative morphology. Consequently, even minor inflectional differences may substantially inflate token-level error rates even when the ASR output and the reference retain considerable semantic overlap.

This study conducted a fine-grained analysis of surface-level differences between human annotations and WhisperX transcriptions, such as spacing, punctuation, and writing conventions. We developed an initial pre-processing pipeline to automatically clean and normalize the transcripts, using the human annotation as the reference standard. This pipeline is currently being further refined to handle a broader range of Korean transcription-normalization issues, including punctuation, spacing, colloquial spellings, interjections, onomatopoeic expressions, non-lexical vocalizations, and numerical expressions. Moreover, we identified two major categories of ASR errors for Korean caregiver-child interactions: (1) *transcription scope mismatches* where the segmented human annotations and the automatic transcriptions differ in what is treated as transcribable. In these cases, background noise, TV or media sounds, and onomatopoeic or non-lexical vocalizations, such as infant vocalizations, may be transcribed differently or omitted entirely; (2) *ASR output errors,* that is, genuine speech recognition failures attributable to the tool, including hallucinated repetitions of infant babbling, phonetic or lexical substitutions, and boilerplate hallucinations. These transcription errors, together with transcription normalization issues, contributed to elevated WER and CER values between human and WhisperX transcriptions. Because the normalization pipeline remains under development, the error rates reported here should be interpreted as preliminary estimates rather than final model-performance scores.

We examined whether lexical and embedding-based semantic similarity metrics could 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 transcripts automatically generated 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 lexical and semantic similarities between the two types of transcripts. We report preliminary results based on the current version of the preprocessing pipeline, which normalizes selected cases of 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 indicate that only an average of 60% of the 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. These values, however, are expected to change as the preprocessing and normalization pipeline is refined.

Our findings suggest that, despite low preliminary surface-form-based scores, embedding-based semantic similarity between human and WhisperX transcripts of Korean caregiver-child conversations may reveal a substantial degree of shared meaning. Nevertheless, the overall semantic similarity score should be supplemented by fine-grained sentence-to-sentence embedding-based analyses in order to obtain a refined map of the semantic match. Moreover, while Sentence-BERT and BERTScore capture meaning according to token embeddings, these metrics do not fully represent the pragmatic dimensions of CDS. Further pragmatic and Korean-culture-sensitive analysis could be conducted using Large Language Models and Chain-of-Thought or other reasoning-oriented prompting methods 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*.