Dialectal Fairness in Korean Speech-to-Text Technology

Keywords: datasets, corpus audit, speech language technologies, dialectal variation, algorithmic fairness

Albeit rising awareness of its importance, fairness in machine-learning based Speech-to-Text (STT) systems still remains an understudied topic, especially so when it comes to speech domains outside of "Standard American English." Thus, when Korean itself is an understudied language in the Natural Language Processing (NLP) and Automatic Speech Recognition (ASR) field, it is natural to assume that Korean dialects are even more understudied. Korean dialects have rarely been in the center of focus in the Korean speech technologies research, with very few prior studies even incorporating the variation. Dialect Identification (DID) may be the only area where Korean dialects are explicitly explored, but even the field lacks literature ([5, 4]. Areas like ASR and Speech Synthesis mostly brush upon the possibility of dialects (ASR [12, 7]; Speech Synthesis [6]). Only one past study investigated the dialectal disparity present in the state-of-the-art Korean speech technologies [9]. While [9] attempted to compare the performance of three different commercial STT API services on different Korean dialects, their experiment was poorly set up, with no clear mention of the number of participants recruited or the number of utterances used, and no clear metric of the evaluation result.

Korean Dialects and Speech Technologies

DataDam Data Audit

DataDam Collection Audit

Call to Action A large part of the challenge in using large speech datasets for training comes from the fact that those large datasets have not been properly audited and controlled for quality, especially so in mitigating whatever underlying bias they may have. The effect that the training datasets have in the resulting technology has been studied before, and many argue the importance of "un-biasing" datasets in order to obtain un-biased results [8, 11, 10]. However, most of the literature focus on socially oriented datasets that are subjectively human-annotated, such as hate-speech, or datasets explicitly using biased language.

Dialect variation is a

The Korean language can be largely grouped into six large dialectal zones based on geographical regions: Northwestern (Pyongan), Northeastern (Hamgyong), Central (Gyeonggi, Hwanghae, Gangwon, Chungcheong), Southwestern (Jeolla), Southeastern (Gyeongsang), and Jeju [1]. Among these, the Northern dialects and a part of the Central dialects (Hwanghae) need to be excluded since they are located in North Korea. The Central dialect can be further divided into each its own, with Gyeonggi dialect being used interchangeably to represent Seoul Standard dialect. Chungcheong, Jeolla and Gyeongsang is administratively divided into South and North provinces, respectively, but they are usually merged in dialectological analyses [4]. Our research also follows the convention of the five dialects, with Gyeonggi dialect representing the "Seoul Standard" speech, which will be referred to as Standard dialect throughout the paper.

In 2020, the Korean Ministry of Science and ICT's Data Dam(https://aihub.or.kr/) initiative released more than 150 types of data built for training AI in a broad spectrum of society. This included five real-life speech datasets of Korean regional dialects, henceforth referred to as the Korean Dialect corpus, adding to a collection of Standard datasets they released previously. Unlike data collected from reading out a written text in laboratory settings that force people to speak in ways significantly different from their actual speech patterns, spontaneous speech spoken between two or more people in naturalistic settings can provide great insight into how people actually speak in real life. Previously having only a handful of Korean speech data large enough for machine learning (ML) training purposes, such a big dataset built specifically to incorporate different dialects provides many opportunities to entities looking to add large dataset to their training corpus. With this opportunity comes the need to audit this dataset, in order to assure that no dialectal disparity occurs in the resulting speech technologies due to a biased dataset.

Present research has two parts to the research:

It has been argued that one main challenge of recognizing dialect features computationally is the lack of labeled data ([2]),

Quantitative Analysis Perplexity is the standard performance metric for language models, and is simply understood as how uncertain a language model is when it generates a new token. Consequently, better language models have lower perplexity, indicating that the language model is not hesitant to generate the next token in the utterance. Although the underlying language model of commercial Korean STT products are unknown, the present research used the most updated, publicly available Korean language model KoGPT2 (https://github.com/SKT-AI/KoGPT2) to calculate the perplexity of each segmented utterance. For all possible utterances, pre-determined by the annotators' transcription, only utterances with more than 10 words were used as input. Gyeongsang has lower perplexity than expected

Gangwon has higher perplexity than expected

Jeju, Jeolla and Chungcheong shows expeceted perplexity

Dialect Density Measure (DDM), as used in the pioneering work of [3], is calculated for 50 randomly selected snippets from each dialect. Specifically, for each snippet, we counted the number of words that expressed the linguistic features of the dialect, including phonology and vocabulary, and then divided by the number of total words in the utterance for normalization.

Qualitative Analysis Consistency in the data collection methodology is of extreme importance (cite), especially so in the transcription of the speech dataset.

Apart from the inter-dataset differences between the Korean Dialect corpus and other datasets provided by the AI HUB, which could be mitigated with different pre-processing methods, intra-dataset differences between the dialects within the Korean Dialect, or even more detrimental, within a single dialect, is more critical to the dataset usage. What we found in the Korean Dialect corpus was:

- Datasets had different annotations for speakers, even within
- Datasets had different number of speakers for each conversation.

Transcription is not the sole inconsistent methodology the Korean Dialect corpus showed evidence of. While the The unexpected result in the high perplexities of Gangwon dataset may also be caused by the inconsistent data collection methods. Gangwon has

inconsistent

References

[1] Lucien Brown and Jaehoon Yeon. Varieties of contemporary korean. In *The Handbook of Korean Linguistics*, pages 459–476. John Wiley & Sons, Inc, May 2015.

[2] Dorottya Demszky, Devyani Sharma, Jonathan H. Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. Learning to recognize dialect features. *CoRR*, abs/2010.12707, 2020.

[3] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

[4] Jooyoung Lee, Kyungwha Kim, and Minhwa Chung. Korean dialect identification based on intonation modeling. In *2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 168–173, 2021.

[5] Jooyoung Lee, Kyungwha Kim, Kyuwhan Lee, and Minhwa Chung. Gender, age, and dialect identification for speaker profiling. 10 2019.

[6] Sungwoo Moon, Sunghyun Kim, and Yong-Hoon Choi. Mist-tacotron: end-to-end emotional speech synthesis using mel-spectrogram image style transfer. *IEEE Access*, 10:25455–25463, 2022.

[7] Kihun Nam. A study on processing of speech recognition korean words. *The Journal of the Convergence on Culture Technology*, 5(4):407–412, 2019.

[8] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[9] Heesu Roh and Kang-Hee Lee. A basic performance evaluation of the speech recognition app of standard language and dialect using google, naver, and daum kakao apis. *Asia-pacific Journal of Multimedia services convergent with Art, Humanities, and Sociology*, 7:819–829, 2017.

[10] Zeerak Waseem. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas, November 2016. Association for Computational Linguistics.

[11] Maximilian Wich, Jan Bauer, and Georg Groh. Impact of politically biased data on hate speech classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 54–64, Online, November 2020. Association for Computational Linguistics.

[12] Hyun Jae Yoo, Sungwoong Seo, Sun Woo Im, and Gwang Yong Gim. The performance evaluation of continuous speech recognition based on korean phonological rules of cloud-based speech recognition open API. *International Journal of Networked and Distributed Computing*, 9(1):10, 2021.