Auditing Korean Speech Datasets for Dialectal Fairness in Speech-to-Text Applications

Anna Seo Gyeong Choi, Allison Koenecke
(sc2359@cornell.edu, koenecke@cornell.edu)
Cornell University

Introduction

- Speech-to-Text (STT) advances are relatively new for the Korean language, especially for the five non-"standard" Korean dialects

- In 2020, the Korean Ministry of Science and ICT's DataDam initiative released a 2.5 TB speech dataset of Korea's five non-SK dialects to encourage practitioners to use as training data for STT models
- In addition to Seoul's "standard" Korean (SK), the five dialectal zones are based on geographic regions: Chungcheong (CC), Gangwon (GW), Jeolla (JL), Gyeongsang (GS), and Jeju (JJ)
- **Research question**: is the novel DataDam dataset sufficient to yield **equitable STT outcomes across Korean dialects?**
- We perform two audits: a qualitative audit of the data collection process, and a quantitative audit of the speech data itself

Qualitative Audit: DataDam Collection

Speech Collection Inconsistencies

- Speech in different dialects were confounded by having different numbers of speakers in conversations
- GW monologues appear to be differently-designed to deliver more dialectal features (consisting of prompt-reading rather than engaging in spontaneous speech)

Transcription Inconsistencies

- Transcription conventions were not released, formatting was inconsistent, and transcriptions lacked basic grammar and spelling checks
- We identified mistranscriptions in 57/500 transcripts selected for review at random

Conclusion

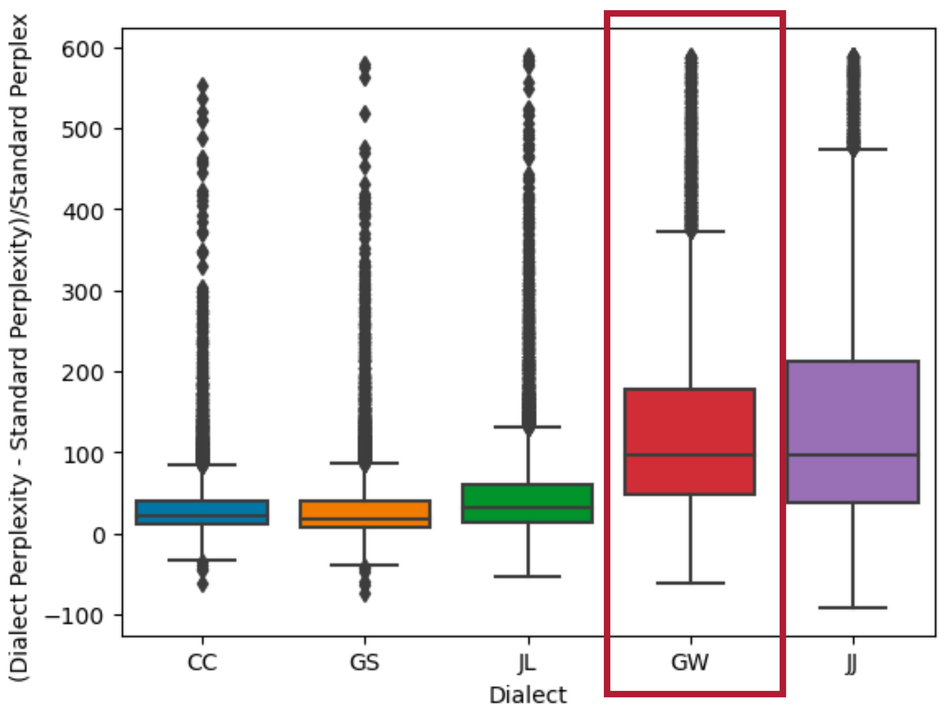
- The DataDam dataset collected on non-standard Korean speech (a**) is not reflective of the true dialectal variation** found across Korea, especially for dialects such as GW; and (b) **contains several technical errors**
- Despite good intentions, **the limited diversity and accuracy** of the DataDam dataset could lead to negative downstream effects, such as (a) lower transcription accuracy for speakers of non-"standard" dialects; and (b) caricaturing linguistic features of dialects
- We hope to raise awareness of the limitations of using the DataDam to train STT models, and advocate for further data collection and cleaning of dialect-dense Korean speech

Quantitative Audit: DataDam Speech Data

We use two metrics (perplexity and DDM) to audit whether the speech data are reflective of the underlying dialects, and find that DataDam may be particularly unrepresentative of the Gangwon (GW) dialect

Perplexity

- Perplexity measures the degree to which a language model is uncertain when generating a new token. We calculate perplexity using KoGPT2 on utterances with > 3 words
- When matched on the same utterance, perplexities for dialectal forms were higher than standard forms (expected behavior if KoGPT2 is trained primarily on SK)
- The ordering of average perplexities is mostly expected: JJ has the highest perplexity, due to the dialect being the most prominently differing from the standard form, while CC has the lowest perplexity, due to the dialect being the least characteristic across dialects.
- However, **GW has surprisingly high perplexity** despite being comparable in dialect features to CC. Meanwhile, **GS and JL have surprisingly low perplexity**



DDM

- **Dialect Density Measure (DDM)** measures dialect "strength" by calculating the share of words in an utterance that are spoken with pre-defined dialectal features
- DDM analyses yield mostly similar results to the perplexity ordering: CC has low DDM in general, and JJ has high DDMs
- Surprisingly, **GW has higher DDMs than CC and JL**, when GW should have relatively few distinct linguistic features