Applying Big Data and Automation Techniques in Phonetics: A Case Study on Hyperarticulation in Korean Word-Initial Stops

Cheonkam Jeong and Andrew Wedel
University of Arizona

The recent availability of large-scale speech data offers a significant opportunity for phonetic research, yet the complexities of data processing have limited its utilization. Advances in deep learning now enable the effective management of these complexities. We introduce strategies for preprocessing and automating phonetic measurements in large-scale natural speech data. Our study leverages big data to examine the impact of lexical competition on hyperarticulation in Korean word-initial aspirated and lenis stops (Jeong & Wedel, 2022) in the context of the ongoing tonogenetic sound change in Korean (Silva, 2006; Bang et al., 2018).

We analyzed the NIKL DIALOGUE 2020 v1.3 dataset (National Institute of Korean Language, 2022) encompassing dialogues from 2,739 speakers, with associated audio files and transcripts. We converted JSON and PCM formats to CSV and WAV, yielding 638,254 utterances. Selecting 106,358 utterances with initial aspirated and lenis stop consonants, we utilized **Pandas** for extraction and KoNLPy's **kkma** parser (Park & Cho, 2014) for minimal pair identification. VOT and F0 intervals were annotated using Dr.VOT (Shrem et al., 2019) and the Montreal-forced aligner (McAuliffe et al., 2017), with pitch extraction via a pretrained model and **Parselmouth**. Batch processing facilitated the efficiency of these tasks.

We validated Dr.VOT's VOT measurements by comparing its output with those of two independent human annotators and Dr.VOT's across 50 WAV files for each laryngeal type. Through Pairwise Intraclass Correlation Coefficient (ICC) calculations and separate *t*-tests, we found high ICC values and no significant *t*-test discrepancies, affirming Dr.VOT's precision. However, it is important to note that this assessment was influenced by the exclusion of low-quality files and those with initial fillers, which were omitted from the error analysis. Moreover, Dr.VOT's default assignment of a 5ms VOT for faint or unclear bursts adds complexity to the analysis. Its limitation to the first syllable poses challenges with filler words, often leading to inflated values until the first vowel.

To address octave halving errors in our dataset, we used a semi-automatic approach summarized as in Algorithm 1. Octave errors were suspected when the pitch of an initial vowel 'f0(v1)' fell below half that of the subsequent pitch 'f0(v2).' In the first round, we analyzed approximately 10% of this set for actual errors and found an adjusted pitch floor that corrected them, applying it to this set. In subsequent rounds of error correction, we excluded the previously corrected set and checked for further octave errors using progressively laxer criteria, e.g. $f0(v1) < 3/4 \times f0(v2)$; $f0(v1) < 3/2 \times f0(v2)$, correcting each time using a manually modified floor. This correction cycle was repeated until errors were absent, completed over three rounds in our case. The sequence of pitch settings adjustments was documented in *settings List* as [1/2, 3/4, 3/2], indicating the progressive refinement over the correction rounds.

After filtering out errors stemming from automatic measurements, including the aligner's failure to detect pitch, and removing outliers, our dataset shrank from 106,358 to 30,901 utterances across 2,024 speakers. Analysis focused on speakers with at least 5 tokens for each laryngeal type, yielding 6,986 tokens. Employing a Bayesian hierarchical model via **brms**, we observed substantial hyperarticulation in VOT (β = 3.91, CI = [2.71, 5.09]) and F0 (β = 0.34, CI = [0.20, 0.48]) for aspirated stops, showing a 7.82 ms increase in VOT and a 0.68 semitone rise in F0 between the first and second vowels in minimal pairs, as illustrated in Figures 1 and 2. Despite the considerable data reduction in automatic processing, our findings underscore big data's viability for detailed phonetic studies. Following error analysis, we've identified potential improvement strategies. Our future endeavors include training our own data and implementing speech enhancement techniques like autoencoder noise reduction. These methods aim to yield more precise measurements of VOT and F0, thereby minimizing substantial data loss.

Algorithm 1 Semi-automatic octave error correction algorithm

Input: Pitch values of each vowel pair: *f0*(v1, v2), List of pitch settings: *settingsList*
Output: Corrected pitch values
for iteration = 1 to maxIterations **do**
  for each set of pitch settings in settingsList **do**
    for each voice sample pair (v1, v2) with detected error **do**
      if f0(v1) ≥ *settingsList*[*currentSettingsIndex*].*threshold* × f0(v2) **then**
        CorrectPitch(v1,v2,settingsList[currentSettingsIndex])
      else
        MarkFileAsCorrected(v1,v2)
      end if
    end for
  end for
end for

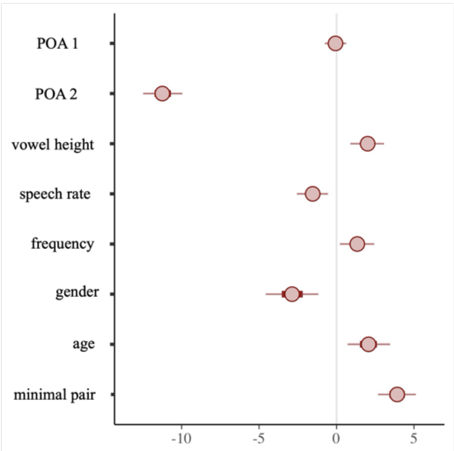

Figure 1. VOT model with aspirated stops

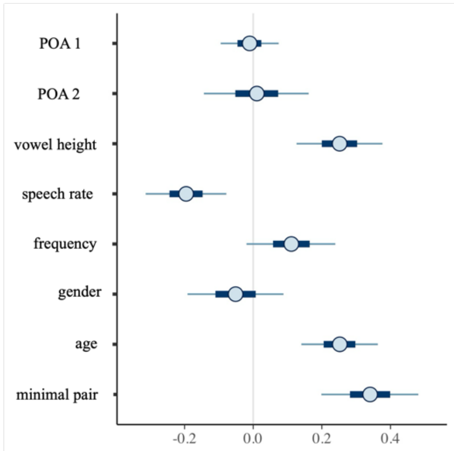

Figure 2. F0 model with aspirated stops

References

Bang, H. Y., Sonderegger, M., Kang, Y., Clayards, M., & Yoon, T. J. (2018). The emergence, progress, and impact of sound change in progress in Seoul Korean: Implications for mechanisms of tonogenesis. Journal of Phonetics, 66, 120-144.

Jeong, C., & Wedel, A. (2022). The Effect of Cue-specific Lexical Competitors on Hyperarticulation of VOT and F0 Contrasts in Korean Stops. In Proceedings of the Annual Meetings on Phonology.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In Proc. Interspeech 2017 (pp. 498-502). doi:10.21437/Interspeech.2017-1386

National Institute of Korean Language. (2022). NIKL Korean Dialogue Corpus (audio) 2020 (Version 1.3). Retrieved from https://corpus.korean.go.kr

Park, E. L., & Cho, S. (2014, October). KoNLPy: Korean natural language processing in Python. In Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology (Vol. 6, pp. 133-136). Chuncheon Korea.

Silva, D. J. (2006). Acoustic evidence for the emergence of tonal contrast in contemporary Korean. Phonology, 23(2), 287-308.