pISSN 2005-8063
eISSN 2586-5854
2023. 12. 31.
Vol.15 No.4
pp. 101-107

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2023.15.4.101

Analyzing vowel variation in Korean dialects using phone recognition*

Jooyoung Lee[1] · Sunhee Kim[2] · Minhwa Chung[1,**]

[1]*Department of Linguistics, Seoul National University, Seoul, Korea*
[2]*Department of French Language Education, Seoul National University, Seoul, Korea*

Abstract

This study aims to propose an automatic method of detecting vowel variation in the Korean dialects of Gyeong-sang and Jeol-la. The method is based on error patterns extracted using phone recognition. Canonical and recognized phone sequences are compared, and statistical analyses distinguish the vowels appearing in both dialects, the dialect-common vowels, and the vowels with high mismatch rates for each dialect. The dialect-common vowels show monophthongization of diphthongs. The vowels unique to the dialects are /we/ to [e] and /ʌ/ to [ɯ] for Gyeong-sang dialect, and /uɥi/ to [ɯ] in Jeol-la dialect. These results corroborate previous dialectology reports regarding phonetic realization of the Korean dialects. The current method provides a possibility of automatic explanation of the dialect patterns.

Keywords: Korean, dialects, vowels, speech recognition

1. Introduction

A dialect of a language is a form of language variety used by a community of speakers with shared social status (Labov, 2006, 1973), ethnic backgrounds (Hasan, 2004; Major et al., 2005), or geographical areas (Clopper & Pisoni, 2006). Typical characteristics of dialects include a shared phonetic inventory among different dialects of a language, and mutual intelligibility among speakers of different dialects (Chambers & Trudgill, 1998).

One main research interest in the field of dialect speech is the discrimination of individual dialects based on acoustic cues. The task of discriminating regional dialects has long been explored by using linguistic approaches, particularly in prosody and phone segments. In prosodic studies, Vicenik & Sundara (2013) conducted perception experiments with adult listeners, examining the use of pitch cues for the perceptual discrimination of English dialects. Rouas (2007) proposed an automatic method for modeling prosodic variations for language and dialect discrimination. This method involves learning prosodic variations through the separation of phrase and accentual components of intonation in various languages and Arabic dialects. Other studies have focused on acoustic phonetics methods, particularly on phonetic segments of speech, for dialect discrimination. McCullough et al. (2019) demonstrated a comparison of vowel formants between four American English dialects. Clopper & Pisoni (2004) designed a set of sentences that included target vowels and proceeded with measurements of

* This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2021-0-00575, "Development of Voice Phishing Prevention Technology Based on Speech and Text Deep Learning").

** mchung@snu.ac.kr, Corresponding author

Received 22 November 2023; Revised 11 December 2023; Accepted 11 December 2023

formant values, examining intended acoustic-phonetic properties.

Research in acoustic phonetics often reviews the phonetic realization of vowel phonemes in the context of dialect discrimination (Chen, 2008; Clopper & Pisoni, 2004; McCullough et al., 2019). The central concept of this approach is that the same vowel phoneme can be pronounced differently across dialects. Studies examining dialect vowels typically involve extracting formants from vowel segments (Chen, 2008; Diehl et al., 1996; Kim et al., 2006). However, formant extraction depends on determining the range of a vowel segment and a time point for calculating formants within the selected segments, which can result in varying values under different conditions. Particularly, diphthongs present a challenge in formant extraction, as measuring formant transitions is difficult when identifying a transition boundary within a diphthong (Park, 2019; Zhao et al., 2023). Moreover, there is ongoing debate about whether diphthongs are considered as one dynamic target or two static targets (Xu et al., 2023).

Another limitation of previous experiment-based dialect studies is the speech style limited to read speech. Since the goal of these studies is to examine different phonetic features of the same vowels from different speaker groups, controlling experiment stimuli to ensure the target vowels are sufficiently pronounced by the speakers is important, and read speech with a given sentence script is effective in collecting the data. However, predefined read speech scripts reduce the likelihood of dialect vowels being pronounced. This is because an orthographic script can unintentionally lead dialect speakers to pronounce words as they are written in the text. In contrast, Spontaneous speech includes dialect-related features as it is naturally spoken. The downside, however, is the difficulty in collecting data of interest; there is no script for spontaneous speech, making it time-consuming to identify the target vowels and extract formant features.

This study proposes application of an automatic approach to detect phonetic realization of the dialectal vowels, both monophthongs and diphthongs, that does not involve the formant measurement of a vowel segment in spontaneous speech environment. Our research questions are as follows:

(1) Is there a way to compare vowels of different dialects without the process of formant extraction?
(2) Is there a way to use spontaneous speech for dialect analysis?

A research field of mispronunciation detection and diagnosis (MDD) in second language education has a long history of using methods that compare pronunciation sequences of native speakers and language learners, by analyzing phone-level mismatch between the phone sequences using a phone recognizer (Li et al., 2016; Lin & Wang, 2022; Yeo et al., 2023). Since the pronunciation comparison is done with output phones from an automatic speech recognition model, the process does not involve the formant extraction. The output vowels, both monophthongs and diphthongs, are effectively examined in mispronunciation detection of non-native speech. The same idea is applicable to the dialect phone analysis, in which phone-level comparison is made between the pronunciations of the standard language and the dialects using a phone recognizer trained on the standard language. Note that despite the use of a speech recognizer, terminology for an error rate of dialect vowel recognition is a “mismatch rate” instead of an “error rate” because a canonical vowel recognized as another vowel in a dialect does not convey any erroneous meaning.

According to Yoon et al. (2015), South Korea has six major dialects: (i) the north-western dialect, (ii) the north-eastern dialect, (iii) the central dialect, (iv) the south-western dialect, (v) the south-eastern dialect and (vi) the Cheju dialect. The current study focuses on the two southern dialects, Gyeong-sang (south-eastern) dialect and Jeol-la (south-western) dialect. Main distinction of Gyeong-sang dialect from other dialects is a vowel merger between /ʌ/ and /ɯ/ (Jang, 2021; Kwak, 2003; Paek, 1999; Park, 2022). Regarding the merging direction of the vowel pair, Kwak (2003) claims the merging direction is towards [ɯ], whereas Jang (2021) claims the opposite direction that the merger occurs toward [ʌ] based on formant values between the two vowels. Also, Park (2022) confirms with experiments that the merger was observed from senior speakers only, and younger generation was able to make phonetic difference between /ʌ/ and /ɯ/. Vowel /y/ is pronounced as either [wi], [i], or [u] in Gyeong-sang dialect depending on speaker age, and Bae (2012) reports speakers in their 60s pronounce /y/ as [u] in Gyeong-sang dialect. A distinctive vowel characteristic found in Jeol-la dialect is pronunciation of diphthong /uɟi/ as [ɯ] (Jang, 2019). Also, vowels /y/ and /ø / are reported to pronounce as [i] and [e], respectively (Kwak, 2003). Park (2003) reports two monophthongs /u/ and /a/ in a consecutive order are pronounced either as a diphthong [wa] or [wo] We proceed to compare the results of the proposed phone comparison method with the reports in the dialectology studies, with a specific focus on the vowel characteristics examined in the literature survey.

The rest of the paper is as follows: Section 2 presents the methodology of the current work, focusing on the process of the phone sequence comparison for dialect vowel analysis. Section 3 reports dialect-common and dialect-unique vowels and supports the effectiveness of the proposed method by comparing the outcomes with the previous dialectology studies. Section 4 summarizes the paper and concludes the work.

2. Methodology

The overview of the experiment process is illustrated in Figure 1. The process consists of a fine-tuned Korean phone recognizer, grapheme-to-phoneme conversion (G2P conversion), and phone sequence pair comparison. Dialect speech samples are fed into the end-to-end Korean phone recognizer and the G2P conversion module to obtain pronounced and canonical phone sequences, respectively. The paired sequence outputs are then compared phone-wise to calculate phone patterns. The phones with high substitution rates are considered as the vowels that are phonetically realized in the dialects.

2.1. Dialect Dataset

Table 1. Distribution of the dialect dataset

Dialects	Num. of speakers	Utterances	Audio size
Gyeong-sang	51	33,076	71 h 10 m
Jeol-la	47	37,261	81 h 41 m

The dialect dataset in this work is AI-Hub Senior Spontaneous Speech corpus (AI-Hub, 2021). This public dataset includes recordings Korean senior speakers in their age ranging from 60s to

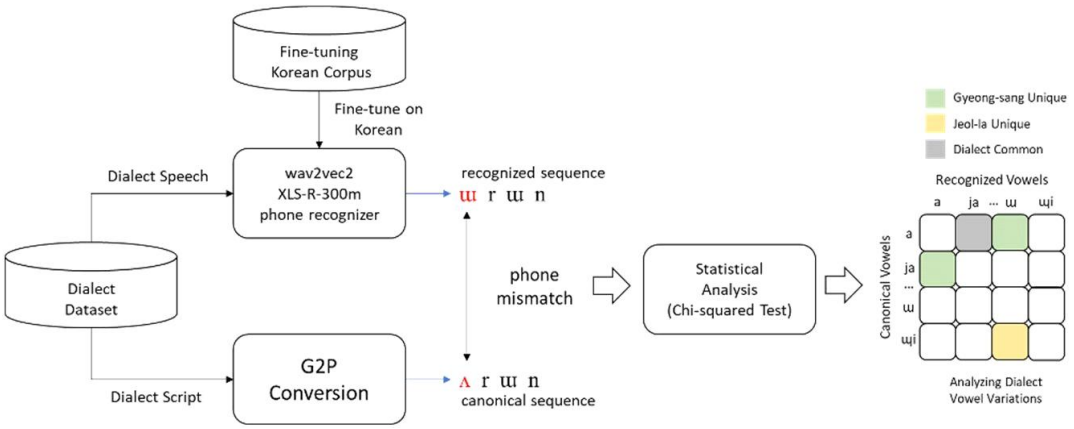

Figure 1. An overview of the vowel analysis of dialects using phone-level comparison.

90s. We use this dataset to obtain dialectal patterns embedded in speech of the senior speakers. Younger speakers in their ages around 20s to 30s, on the other hand, tend to show less dialectal characteristics due to the influence of mass media and public education. Another reason to use this dataset is spontaneous speech which is assumed to include dialectal vowels in a naturally speaking manner. Also, a previous work regarding the speaker age and dialectal speech supports the use of speech from the senior speakers (Park, 2022).

This work uses a part of the dataset that is labeled with Gyeong-sang and Jeol-la dialects. Out of 365,237 utterances in total, Gyeong-sang and Jeol-la dialects take up 9.06% and 10.20% of the entire dataset, respectively. Table 1 summarizes the data distribution of the Korean dialects. It can be seen from the distribution that both dialects have around 50 speakers with the audio size of 71 hours and 81 hours, respectively. We believe the data size of the dialects is enough to analyze the vowels from the recognized phone sequences.

2.2. Korean Phone Recognizer

A phone recognizer is an automatic speech recognition (ASR) model that transcribes in phonetic transcription given an audio utterance. Speech recognition technology has achieved significant improvement in recognition performance with recent end-to-end approaches (Baevski et al., 2020), which considers only the input audio and the output target text sequence without intermediate pronunciation models. A phone recognizer prints out what is phonetically spoken; thus, the resulting phone sequence is used as a pronounced sequence that includes dialectal vowels.

In this study, an end-to-end Korean phone recognizer is trained by fine-tuning a pre-trained speech model with a Korean read speech corpus that is designed to be phonetically balanced. Fine-tuning is a process of adapting a pre-trained model to a specific domain. Instead of training a model from scratch, fine-tuning uses a model that is trained with very large datasets with general purposes. Fine-tuning process allows a model to show better and more robust performance on a target task even with a small set of domain specific data compared to when training a model from scratch. The pre-trained module used is wav2vec 2.0 XLS-R (Babu et al., 2022) with 300 million parameters (300 m), which is a large-scale model for cross-lingual representation trained with audio samples in 128 languages. By taking advantage of the large range of languages and audio information trained in the pre-training phase, the end-to-end Korean phone recognizer enables accurate clustering of individual Korean phonemes, especially the vowels. The fine-tuned phone recognizer emits dialectal phone sequences (i.e. [ɯ r ɯ n]) described in Figure 1.

We use a dataset different from the AI-Hub dialect dataset for fine-tuning an end-to-end Korean phone recognizer. The dataset for fine-tuning is an in-house Korean read speech corpus that is phonetically balanced, and consists of 60,000 utterances from 600 speakers (100 utterances per speaker) with the audio size of 120 hours in total. 54,000 utterances (540 speakers; 108 hours) are used for fine-tuning the pre-trained XLS-R (Babu et al., 2022) (300 m) model, and the remaining 6,000 utterances (60 speakers; 12 hours) are used for evaluating performance of the fine-tuned model.

Before applying the fine-tuned model to phone recognition of the dialect speech, it is important to examine the performance of the fine-tuned phone recognizer. The model performance achieves 3.88% in phone error rate, which is low enough to be used for finding the dialectal vowels. Despite the reliable overall performance, the model needs further investigation on recognition errors on the vowel-level. This is because if a high error rate is observed in certain vowels, validity of the analysis on the dialects using the phone recognizer is weakened.

Figure 2 shows a confusion matrix of the canonical and the recognized vowels. The canonical vowels on the y-axis indicate the reference vowels that are expected as the standard pronunciation. The recognized vowels on the x-axis mean the output of vowels of the recognizer that is actually spoken in an utterance. The "*" symbol denotes an empty symbol used to indicate insertions and deletions of the recognition. Substitution from "*" to a vowel phone is interpreted as an insertion, while substitution from a vowel to "*" indicate a deletion. The number in each cell of the confusion matrices is a normalized value by dividing the frequency value of the cell by the total frequency value of the canonical vowel. Since the fine-tuning dataset does not include dialectal accents, the recognizer is expected to show low error rates in all vowels.

The recognition performance in most of the vowels achieves accuracy above 90%, yet some diphthongs (/je, wi, ɯi/) show a relatively low phone recognition rate below 90%. /je/ and /wi/ are misrecognized as [e] and [i], respectively, indicating monophthongized recognition errors. /ɯi/ is substituted to [e], and this is assumed to be pronunciation of [e] of /ɯi/ positioned at word-final as a genitive case. Insertion error patterns are observed in monophthongs /a, e, i,

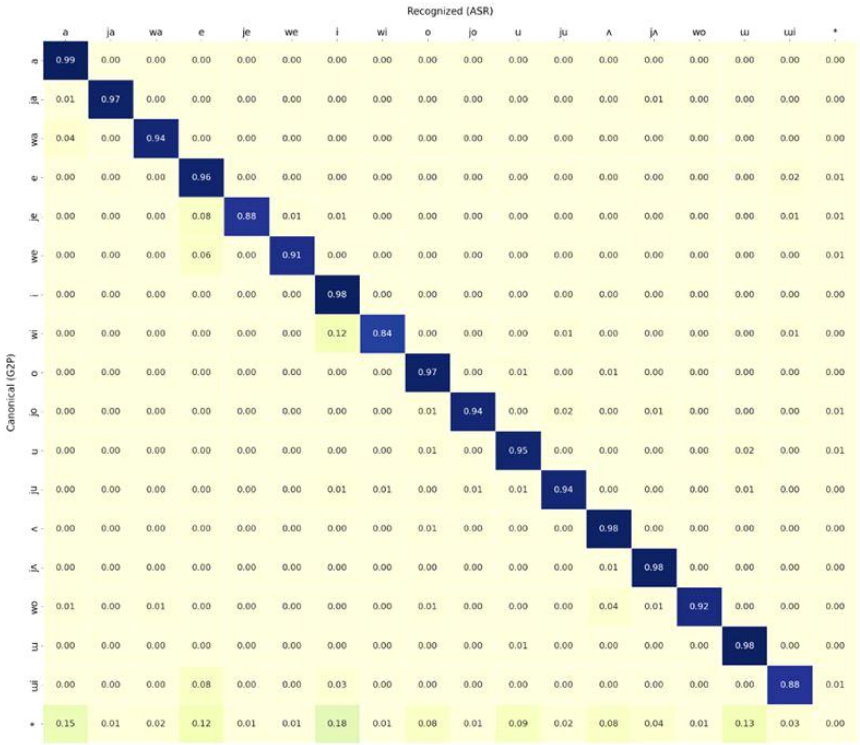

Figure 2. Confusion matrix of vowels from the Korean phone recognizer.

o, u, ʌ, ɯ/. This may come from a higher frequency than the diphthongs. Deletion rates are relatively very low in all vowels with 1% or less than 1%. To summarize, the substitutions of /je, wi, ɯi/ to [e, i, e], respectively, and the insertions of the monophthongs from the phone recognition result are excluded in the dialect analysis due to the relatively high error rates.

2.3. Grapheme-to-Phoneme Conversion

Grapheme-to-phoneme (G2P) conversion is a process of expressing pronunciation of a word in predefined phonemic symbols (phonemes) given its written forms (graphemes) (Bisani & Ney, 2008). The G2P conversion enables the representation of a canonical phone sequence, which can be interpreted as "how it is pronounced in the standard language." Thus, the phonemic output of the G2P conversion can be used as a reference sequence to be compared with the recognized pronunciation sequence.

The G2P conversion is processed using Montreal Forced Aligner (McAuliffe et al., 2017) (MFA), an open-source toolkit. The MFA toolkit supports acoustic models of various languages including Korean. The Korean G2P model is trained on 48,682 words with 55 graphemes and 56 phonemes, and shows 0.64% phone error rate and 3.90% word error rate on 5,409 evaluated words, which is sufficiently low to use for the current study. The Korean G2P model allows a user to choose from selection of Hangeul inputs either in a compound form (i.e., 달) or in a sequence of "jamo" forms (i.e., ㄷㅏㄹ). The current work chooses the jamo input form since the G2P output is empirically much more accurate when using the jamo input than the compound form input. Table 2 shows examples of the Korean G2P conversion. The orthographic transcription in the compound form is first processed to be split into the jamo forms using a jamo separation script, which are then used as an input to the G2P conversion.

Table 2. Examples of Korean G2P conversion

Compound form	Jamo form	G2P output
달	ㄷㅏㄹ	t a ɭ
빙수	ㅂㅣㅇㅅㅜ	p i ŋ sʰ u
저금	ㅈㅓㄱㅡㅁ	tɕ ʌ g ɯ m

2.4. Comparison of Phone Sequences

Comparison is made between the canonical phone sequences from the G2P output and the annotated pronunciation sequence from the phone recognition. The Levenshtein distance is calculated to check the number of matched phones and mismatched phones among all vowels from both the canonical and the recognized phone sequences.

2.5. Statistical Analysis

Statistical analysis is proceeded to confirm vowels commonly appearing across the two dialects and vowels that are specifically unique to individual dialects. A chi-squared test of independence is used to evaluate significance of difference between the dialects or the recognition results (the standard language). A 2×2 contingency table consists of two dialects (Gyeong-sang and Jeol-la) or the phone recognizer as an independent variable and two vowel pairs (a canonical vowel recognized as another vowel and one recognized as the same vowel) as a dependent variable. The chi-squared test is conducted on all vowel pairs. To extract dialect vowels that show significant difference from the standard language, the statistical test is conducted on each dialect and the phone recognizer. If both dialects show significant difference from the recognizer and also the mismatch rates of both dialects are higher than the recognizer, the vowel pairs are interpreted as dialect-common. As for the vowels

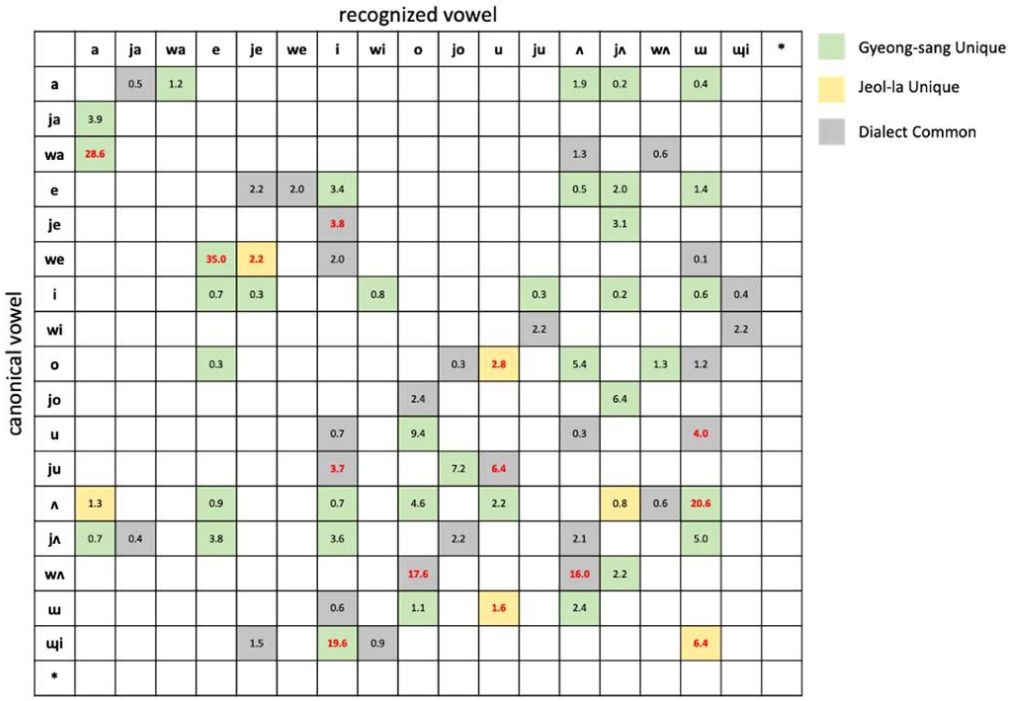

Figure 3. Confusion matrix of dialect-common (gray), Gyeong-sang unique (green), and Jeol-la unique (yellow).

that are unique to individual dialect, the same statistical test is done between the two dialects for the vowels that are. If there is significant difference and the mismatch of one dialect is higher than the other, we interpret that the vowel is unique to the dialect with the higher mismatch rate, hence dialect-unique. The p-value used to determine the significance is set to 0.001 for strict confidence. Also, the observed sample size is limited to five or more since the chi-squared test does not function properly with a sample size less than five.

3. Results & Discussion

This section reports vowels that are shown in both dialects (dialect-common) and vowels that appear exclusively in individual dialects (dialect-unique) based on the statistical results. Figure 3 illustrates a summarized confusion matrix marked with different cell colors for dialect-common and dialect-unique vowels. Mismatch rates in percentage are also shown in the colored cells. The mismatch rates of the dialect-common vowel (gray) are an average of the two rates of the dialects. Also, as stated in section 3, the substitutions of /je, wi, ɯi/ to [e, i, e], respectively, and the insertions of the monophthongs are excluded from the results. This is because the errors from the recognizer may have been influenced from the phone recognizer itself, rather than the characteristics of the dialects. Also, an overall mismatch rate of the dialect dataset is 15.75% from 3,150,822 phone tokens, and mismatch rates of Gyeong-sang and Jeol-la dialects are 18.27% from 1,448,434 tokens and 13.61% from 1,702,388 tokens, respectively.

3.1. Dialect-Common Vowels

The cells colored in gray in Figure 3 indicates mismatched vowels that appear in both dialects. Among all the dialect-common vowels, six vowel pairs with the highest mismatch rates are presented in Table 3. The mismatch rates of the six vowel pairs are also emphasized in red bold in Figure 3.

From the results on the dialect-common vowels in Table 3, except for /u/ recognized as [ɯ], monophthongization of the diphthongs /wʌ, ju, je/ is observed. Also, the canonical and recognized vowels of the four highest mismatch rates are in close relation in terms of place of articulation. /wʌ/ is recognized either as [o] or [ʌ], both of which are closely or identically articulated to monophthong /ʌ/. /ju/ is monophthongized as [u]. /u/ and [ɯ] are also pronounced in similar in place of articulation.

Table 3. Dialect-common vowels

Canonical (G2P)	Recognized (ASR)	Mismatch rate (%)
wʌ	o	17.6
wʌ	ʌ	16.0
ju	u	6.4
u	ɯ	4.0
je	i	3.8
ju	i	3.7

3.2. Dialect-Unique Vowels

Table 4 shows four vowel pairs with the highest mismatch rates of Gyeong-sang and Jeol-la dialects, respectively. The mismatch rates are shown in red bold in Figure 3. As some of the vowel pairs in Table 4 are analyzed in the following subsections, it is confirmed that the resulting vowel pairs are in line with previous dialectology studies, proving the effectiveness of the proposed ASR-based phonetic comparison method to demonstrate dialectal vowels in spontaneous speech without measuring formants.

Table 4. Dialect-unique vowels

Dialect	Canonical (G2P)	Recognized (ASR)	Mismatch rate (%)
Gyeong-sang	we	e	35.0
	wa	a	28.6
	ʌ	ɯ	20.6
	ɯi	i	19.6
Jeo-la	ɯi	ɯ	6.4
	o	u	2.8
	we	je	2.2
	ɯ	u	1.6

3.2.1. Gyeong-sang dialect

From the Gyeong-sang dialect vowel pairs, two of the four vowel pairs are reported in the dialectology studies as dialectal characteristics of Gyeong-sang vowels. The canonical /we/ recognized as [e] is in line with Paek (1999), which claims /we/ in Northern Gyeong-sang dialect has changed to /e/. As for the /ʌ/-[ɯ] vowel pair, there have been many studies regarding a vowel merger between /ʌ/ and /ɯ/ in Gyeong-sang dialect (Jang, 2021; Kwak, 2003; Paek, 1999; Park, 2022). One interesting point to note here is that, judging from the mismatch rates, the merging direction is only towards [ɯ] (mismatch rate of 20.6%) and not towards [ʌ] (mismatch rate of 2.4%), which we can report that the data-driven result confirms the vowel merger occurs in the same direction that Kwak (2003) claimed. There is no clear report regarding /wa/ pronounced as [a] and /ɯi/ pronounced as [i], but with the high mismatch rates we claim they are vowels unique to Gyeong-sang dialect. The /ɯi/-[i] pair, in particular, needs further investigation as the diphthong /ɯi/ is known to have various pronunciation depending on preceding consonants and a position within a word.

3.2.2. Jeol-la dialect

Four pronounced vowels unique to Jeol-la dialect are also shown in Table 4. Jang (2019) states that a diphthong /ɯi/ being pronounced as [ɯ] is a distinctive characteristic in Jeol-la dialect, which is also observed in the current study with the highest mismatch rate among the Jeol-la unique vowel pairs. It is interesting to note that the canonical /ɯi/ appear in both dialects, yet the recognized (pronounced) vowels are different; Gyeong-sang /ɯi/ is pronounced as the second part of the diphthong [i], whereas Jeol-la /ɯi/ is pronounced as the first part of the diphthong [ɯ]. We assume this may be an important cue to distinguish between Gyeong-sang and Jeol-la dialect. The other three vowel pairs have not been clearly dealt with in previous studies, and the current study also cannot easily jump to conclusion that the three pairs are dialect-unique because the mismatch rates are low (1~2%) despite the mismatch rates being statistically higher than Gyeong-sang dialect or the phone recognizer. Also, as Park (2003) reports in the work that two monophthongs /u/ and /a/ in a consecutive order are pronounced either as a diphthong [wa] or [wo], further examination needs to be taken to check if the vowel pairs /o/-[u] and /ɯ/-[u] may have been a part of consecutive monophthongs being pronounced as another monophthong series or diphthongs. This result triggers an extension of research in the future.

4. Conclusion

This paper proposed application of an automatic approach to detect phonetic realization of Korean dialectal vowels, both monophthongs and diphthongs, that does not involve formant measurement of a vowel segment in spontaneous speech environment. The current study focused on analyzing vowel variations of Korean dialects, Gyeong-sang and Jeol-la dialects, by leveraging on wav2vec 2.0-based Korean phone recognition for phone-level mismatch between phone sequences of recognized pronunciation of the dialects and canonical pronunciation of the standard language. Statistical analysis shows characteristics of vowel variations across the dialects and within individual dialects. It is observed that diphthongs are pronounced as monophthongs across the dialects, and pronounced vowels are close to canonical vowels in terms of place of articulation. As for vowel variations in individual dialects, Gyeong-sang dialect poses vowel variations of /we/-[e] and /ʌ/-[ɯ] corroborate previous dialectology reports, and Jeol-la dialect poses a vowel variation of /ɯi/-[ɯ], which is also in line with previous studies. The proposed method presents an effective way to observe characteristics of dialectal vowel variations and analyzes in spontaneous speech environment without formant extraction process.

The current work has limitations that it fails to capture diphthongized merger of two consecutive monophthongs (i.e., /u a/ to [wa]) due to one-to-one phone mapping between the two phone sequences. Also, consideration of surrounding phonemes of the vowels, especially /ɯi/, regarding preceding consonants and a position within a word is absent. Moreover, since the Korean phone recognizer is fine-tuend on the standard language, phone recognition on the dialect speech is influenced by acoustic context from the trained standard language. Thus, the dialect vowel variations from vowel mismatch is not independent from the effect of the standard language and may have resulted in less observation than expected. Future work aims to examine the monophthong-diphthong mapping problem of the proposed phone comparison method, and analyze pronunciation environment of the vowels in terms of onset consonants and a position of the vowels inside a word; whether it is positioned at word-initial or not. Also, preprocessing of comparing models that are trained on either the standard language or the dialects is required to reduce influence of the trained standard language on dialects.

References

AI-Hub. (2021). AI-Hub senior spontaneous speech corpus. Retrieved from https://www.aihub.or.kr

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., ... Auli, M. (2022, September). XLS-R: Self-supervised cross-lingual speech representation learning at scale. *Proceedings of Interspeech 2022* (pp. 2278-2282). Incheon, Korea.

Bae, H. (2012). A study on the aspects of vowel 'ㅟ' change at Daegu. *Korean Language and Literature Society*, *116*, 27-50.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449-12460.

Bisani, M., & Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, *50*(5), 434-451.

Chambers, J. K., & Trudgill, P. (1998). *Dialectology*. Cambridge, UK: Cambridge University Press.

Chen, Y. (2008). The acoustic realization of vowels of Shanghai

Chinese. *Journal of Phonetics*, *36*(4), 629-648.
Clopper, C. G., & Pisoni, D. B. (2004). Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics*, *32*(1), 111-140.
Clopper, C. G., & Pisoni, D. B. (2006). The nationwide speech project: A new corpus of American English dialects. *Speech Communication*, *48*(6), 633-644.
Diehl, R. L., Lindblom, B., Hoemeke, K. A., & Fahey, R. P. (1996). On explaining certain male-female differences in the phonetic realization of vowel categories. *Journal of Phonetics*, *24*(2), 187-208.
Hasan, R. (2004). Code, register and social dialect. *Class, Codes and Control*, *2*, 253-292.
Jang, S. (2019). A study on the orthography of the Jeollabuk-do dialect dictionary. *Korean Language and Literature*, *71*, 97-120.
Jang, S. Y. (2021). Influence of standard Korean and Gyeongsang regional dialect on the pronunciation of English vowels. *Phonetics and Speech Sciences*, *13*(4), 1-7.
Kim, H. G., Choi, Y. S., & Kim, D. S. (2006). An experimental study of Korean dialectal speech. *Speech Sciences*, *13*(3), 49-65.
Kwak, C. G. (2003). The vowel system of contemporary Korean and direction of change. *Journal of Korean Linguistics*, *41*, 59-91.
Labov, W. (1973). *Sociolinguistic patterns (conduct and communication, 4)*. Philadelphia, PA: University of Pennsylvania Press.
Labov, W. (2006). *The social stratification of English in New York city*. Cambridge, UK: Cambridge University Press.
Li, K., Qian, X., & Meng, H. (2016). Mispronunciation detection and diagnosis in L2 English speech using multidistribution deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(1), 193-207.
Lin, B., & Wang, L. (2022, May). Phoneme mispronunciation detection by jointly learning to align. *Proceeding of the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6822-6826). Singapore, Singapore.
Major, R. C., Fitzmaurice, S. M., Bunta, F., & Balasubramanian, C. (2005). Testing the effects of regional, ethnic, and international dialects of English on listening comprehension. *Language Learning*, *55*(1), 37-69.
McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017, August). Montreal forced aligner: Trainable text-speech alignment using kaldi. *Proceedings of Interspeech 2017* (pp. 498-502). Stockholm, Sweden.
McCullough, E. A., Clopper, C. G., & Wagner, L. (2019). Regional dialect perception across the lifespan: Identification and discrimination. *Language and Speech*, *62*(1), 115-136.
Paek, D. (1999). Diachronic changes in Yeongnam dialect. *Journal of Korean Cultural Studies*, *20*, 23-79.
Park, J. (2003). The opacity of vowel harmony in Jeon-Buk dialect. *Korean Language and Literature*, *134*, 155-171.
Park, J. (2022). The phonological perception and articulation by generation in Daegu dialect /ㅡ/ and /ㅓ/. *The Journal of Studies in Language*, *38*(2), 127-142.
Park, S. (2019). On the phonological representation and acoustic properties of the diphthong /i/ in Korean. *Studies in Modern Grammar*, (102), 165-183.
Rouas, J. L. (2007). Automatic prosodic variations modeling for language and dialect discrimination. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(6), 1904-1911.
Vicenik, C., & Sundara, M. (2013). The role of intonation in language and dialect discrimination by adults. *Journal of Phonetics*, *41*(5), 297-306.
Xu, A., Gerazov, B., van Niekerk, D., Krug, P. K., Prom-on, S., Birkholz, P., & Xu, Y. (2023, August). Computational models for articulatory learning of English diphthongs: One dynamic target vs. two static targets. *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 4140-4144). Prague, Czech Republic.
Yeo, E. J., Ryu, H., Lee, J., Kim, S., & Chung, M. (2023, August). Comparison of L2 Korean pronunciation error patterns from five L1 backgrounds by using automatic phonetic transcription. *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 2720-2724). Prague, Czech Republic.
Yoon, T. J., Kang, Y., Han, S., Maeng, H. S., Lee, J., & Kim, K. (2015, August). A corpus-based approach to dialectal variation in Korean vowels. *Proceedings of the 18th International Congress of Phonetic Sciences* (ICPhS 2015). Glasgow, UK.
Zhao, D., Park, J., & Seong, C. (2023). Acoustic characteristics of glides and nuclear vowels of Korean Diphthongs in coarticulation condition. *Han-Geul*, *84*(1), 5-43.

• **Jooyoung Lee**
Ph.D. Candidate, Dept. of Linguistics
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
Tel: +82-2-880-9039
Email: excalibur12@snu.ac.kr
Fields of interest: Korean dialects, ASR, Spoken language processing

• **Sunhee Kim**
Professor, Dept. of French Language Education
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
Tel: +82-2-880-7693
Email: sunhkim@snu.ac.kr
Fields of interest: French phonetics, Spoken language processing

• **Minhwa Chung,** Corresponding author
Professor, Dept. of Linguistics
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
Tel: +82-2-880-9195
Email: mchung@snu.ac.kr
Fields of interest: ASR, Spoken language processing, Computer assisted language learning