COMPARISON OF L2 KOREAN PRONUNCIATION ERROR PATTERNS FROM FIVE L1 BACKGROUNDS BY USING AUTOMATIC PHONETIC TRANSCRIPTION

Eun Jung Yeo[*], Hyungshin Ryu[*], Jooyoung Lee, Sunhee Kim, Minhwa Chung

Seoul National University
{ej.yeo, rhss10, excalibur12, sunhkim, mchung}@snu.ac.kr

ABSTRACT

This paper presents a large-scale analysis of L2 Korean pronunciation error patterns from five different language backgrounds, Chinese, Vietnamese, Japanese, Thai, and English, by using automatic phonetic transcription. For the analysis, confusion matrices are generated for each L1, by aligning canonical phone sequences and automatically transcribed phone sequences obtained from fine-tuned Wav2Vec2 XLS-R phone recognizer. Each value in the confusion matrices is compared to capture frequent common error patterns and to specify patterns unique to a certain language background. Using the Foreign Speakers' Voice Data of Korean for Artificial Intelligence Learning dataset, common error pattern types are found to be (1) substitutions of aspirated or tense consonants with plain consonants, (2) deletions of syllable-final consonants, and (3) substitutions of diphthongs with monophthongs. On the other hand, thirty-nine patterns including (1) syllable-final /l/ substitutions with /n/ for Vietnamese and (2) /ɯ/ insertions for Japanese are discovered as language-dependent.

Keywords: Comparative analysis, Pronunciation error patterns, Automatic phonetic transcription, L1 influence, L2 Korean speech

1. INTRODUCTION

Computer Assisted Pronunciation Training (CAPT) has emerged as an effective tool for non-native speakers, offering cost-effective feedback while overcoming the time and location constraints of traditional language learning [1, 2]. With the advancement of deep learning, novel architectures have been proposed to improve the detection and diagnosis performance of the CAPT system [2, 3, 4]. Linguistic studies were conducted with a different point of view, where the characteristics of non-native speech are explored [5, 6, 7, 8, 9, 10, 11].

The influence of the first language (L1) on the second language (L2) pronunciation has long been investigated. However, due to the requirement of time- and labor-intensive human transcription, analyses were often limited to certain acoustic phenomena from a small number of participants. For example, 127 obstruent-initial words were collected from 22 English-speaking learners to analyze their realizations of Korean stops [5], while 11 words with monophthongs or diphthongs were recorded from 53 Egyptians to understand their Korean vowel perception and production [6]. On the other hand, some studies have taken advantage of deep learning methods including forced alignments [7, 8] or feature extraction [9, 10, 11], which allowed an analysis of larger materials or participants. Nevertheless, their analyses were often limited to a small subset of phone inventories, such as fricatives [7], nasals [9], and vowels [8, 9, 10, 11].

To mitigate the limitation of human transcriptions especially on large-scale analysis, utilizing automatic transcription can be one possible option. Model-based transcriptions have advantages in terms of time- and cost-efficiency, consistency, and objectivity [12]. With self-supervised-based speech models (SSL) showing state-of-the-art performances in various speech tasks including automatic speech recognition [13, 14], this work proposes to employ automatic phonetic transcription for L2 mispronunciation analysis.

This study conducts a large-scale analysis on L2 Korean pronunciation error patterns of speakers from five different L1 backgrounds, Chinese, Vietnamese, Japanese, Thai, and English. Confusion matrices are generated by aligning the canonical and automatically transcribed phones obtained from fine-tuned SSL Wav2Vec2 XLS-R recognizer [13]. The error patterns for each non-native speech are inferred from the confusion matrix, and are scrutinized with respect to L1 background. We expect our analysis to bring new insights into L1-influenced pronunciations and provide linguistic clarity in developing the CAPT system.

[*] Equal contributions.

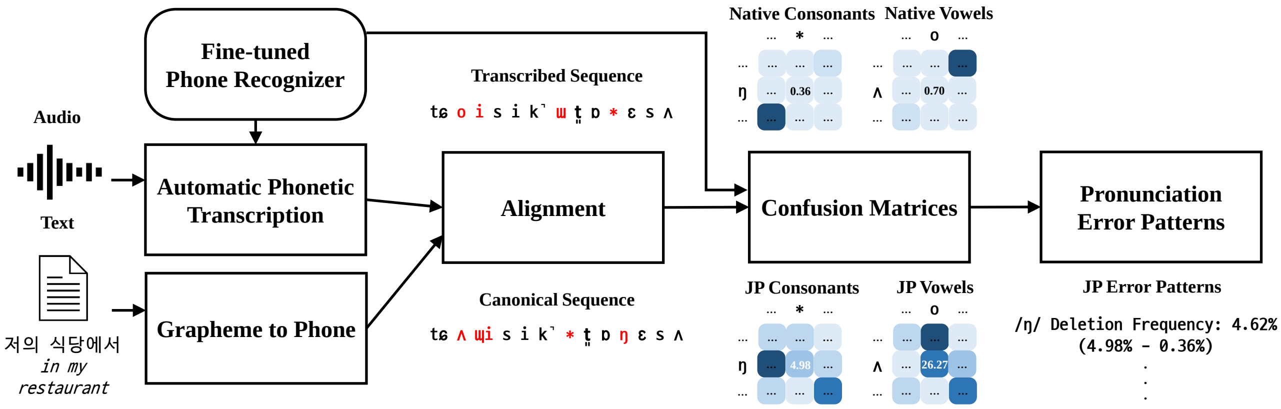


Figure 1: Overview of the pronunciation error patterns analysis using automatic phonetic transcription

2. METHOD

2.1. Corpus

Foreign Speakers' Voice Data of Korean for Artificial Intelligence Learning [15] is an open-source dataset designed to train automatic speech recognition systems for Korean learners. The dataset consists of L2 Korean speech from 1911 speakers with 80 different L1s, where each speaker recorded both read-speech and spontaneous speech. In this paper, we restricted the analysis to five L1 backgrounds, where the most abundant recordings were collected - Chinese (ZH), Vietnamese (VI), Japanese (JP), Thai (TA), and English (EN). In addition, we examined only the read-speech to exclude the influence of speech fluency and focus on pronunciation error patterns. Participants without official scores on the Test of Proficiency in Korean (TOPIK) are also excluded from the analysis. As Table 1 presents, the dataset includes beginners, intermediate and advanced learners for all five language backgrounds.

Table 1: Number of speakers and utterances

Language	Beginner		Intermediate		Advanced		Total Hrs
	spk	utt	spk	utt	spk	utt	
Chinese	2	513	56	18599	291	127811	487.41 h
Vietnamese	12	2837	85	36100	144	74921	412.91 h
Japanese	3	1474	27	12004	140	87915	326.35 h
Thai	44	19773	57	26377	44	26688	260.57 h
English	14	938	15	1682	12	4272	24.34 h

2.2. Analysis

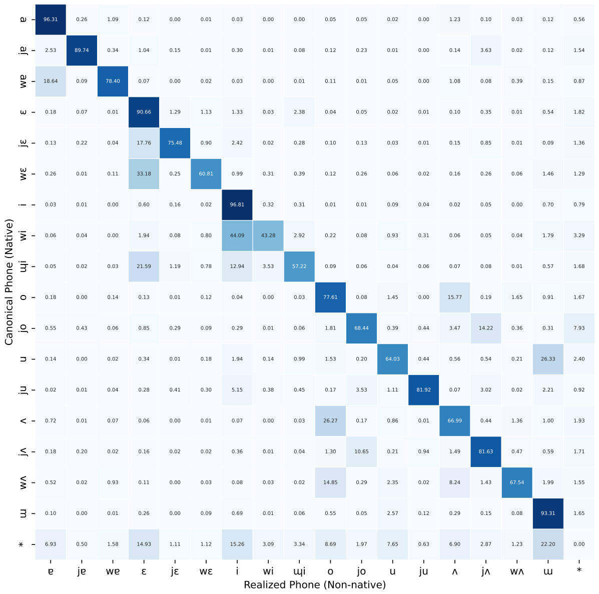
Figure 1 is the overview of our analysis. First, **(1) automatic phonetic transcription** on non-native speech is performed by using the fine-tuned Wav2Vec2 XLS-R phone recognizer. Next, **(2) confusion matrices** for both native and non-native speech are created by aligning the canonical phone sequences and automatically transcribed phone sequences. Finally, **(3) pronunciation error patterns** for non-native speech are extracted from confusion matrices and analyzed regarding five L1 backgrounds. We release the source code and the fine-tuned model for the ease of reproduction.[1]

2.2.1. *Automatic phonetic transcription*

For automatic phonetic transcription, we choose to utilize Wav2Vec2 XLS-R-300m model, which is pre-trained on 128 different languages including Chinese, Vietnamese, Japanese, Thai, and English [13]. The model is fine-tuned with a private, phonetically balanced native Korean read-speech corpus, with a split of 108 hours of the training set (54000 audio files) and 12 hours of the test set (6000 audio files). The transcription used for fine-tuning is generated by trained linguists with the phone set introduced in [16]. Thus the same set is used for error pattern analysis in this study. The phone set includes 40 phones, consisting of 19 consonants, 17 vowels, and 4 allophones. As for allophones, [p̚],[t̚],[k̚] are stops at codas and [ɾ] is /l/ in syllable-initial position. Our model showed a Phone Error Rate (PER) of 3.88% on the test set, which verifies the model's performance.

2.2.2. *Confusion matrices*

Confusion matrices are created for the native dataset, the non-native dataset as a whole, and each L1 group. Canonical phone sequences are generated with the grapheme to phone module in Montreal Forced Aligner [16] and are aligned with the decoded phone sequences from the fine-tuned model, by using SCTK package offered at Kaldi [17]. Figure 2 demonstrates the vowel confusion matrix for Japanese L2 Korean speech which refers



	ɐ	jɐ	wɐ	ɛ	jɛ	wɛ	i	wi	ɰi	o	jo	u	ju	ʌ	jʌ	wʌ	ɯ	*
ɐ	96.31	0.26	1.09	0.12	0.00	0.01	0.03	0.00	0.00	0.05	0.05	0.02	0.00	1.23	0.10	0.03	0.12	0.56
jɐ	2.53	89.74	0.34	1.04	0.15	0.01	0.30	0.01	0.08	0.12	0.23	0.01	0.00	0.14	3.63	0.02	0.12	1.54
wɐ	18.64	0.09	78.40	0.07	0.00	0.02	0.03	0.00	0.01	0.11	0.01	0.05	0.00	1.08	0.08	0.39	0.15	0.87
ɛ	0.18	0.07	0.01	90.66	1.29	1.13	1.33	0.03	2.38	0.04	0.05	0.02	0.01	0.10	0.35	0.01	0.54	1.82
jɛ	0.13	0.22	0.04	17.76	75.48	0.90	2.42	0.02	0.28	0.10	0.13	0.03	0.01	0.15	0.85	0.01	0.09	1.36
wɛ	0.26	0.01	0.11	33.18	0.25	60.81	0.99	0.31	0.39	0.12	0.26	0.06	0.02	0.16	0.26	0.06	1.46	1.29
i	0.03	0.01	0.00	0.60	0.16	0.02	96.81	0.32	0.31	0.01	0.01	0.09	0.04	0.02	0.05	0.00	0.70	0.79
wi	0.06	0.04	0.00	1.94	0.08	0.80	44.09	43.28	2.92	0.22	0.08	0.93	0.31	0.06	0.05	0.04	1.79	3.29
ɰi	0.05	0.02	0.03	21.59	1.19	0.78	12.94	3.53	57.22	0.09	0.06	0.04	0.06	0.07	0.08	0.01	0.57	1.68
o	0.18	0.00	0.14	0.13	0.01	0.12	0.04	0.00	0.03	77.61	0.08	1.45	0.00	15.77	0.19	1.65	0.91	1.67
jo	0.55	0.43	0.06	0.85	0.29	0.09	0.29	0.01	0.06	1.81	68.44	0.39	0.44	3.47	14.22	0.36	0.31	7.93
u	0.14	0.00	0.02	0.34	0.01	0.18	1.94	0.14	0.99	1.53	0.20	64.03	0.44	0.56	0.54	0.21	26.33	2.40
ju	0.02	0.01	0.04	0.28	0.41	0.30	5.15	0.38	0.45	0.17	3.53	1.11	81.92	0.07	3.02	0.02	2.21	0.92
ʌ	0.72	0.01	0.07	0.06	0.00	0.01	0.07	0.00	0.03	26.27	0.17	0.86	0.01	66.99	0.44	1.36	1.00	1.93
jʌ	0.18	0.20	0.02	0.16	0.02	0.02	0.36	0.01	0.04	1.30	10.65	0.21	0.94	1.49	81.63	0.47	0.59	1.71
wʌ	0.52	0.02	0.93	0.11	0.00	0.03	0.08	0.03	0.02	14.85	0.29	2.35	0.02	8.24	1.43	67.54	1.99	1.55
ɯ	0.10	0.00	0.01	0.26	0.00	0.09	0.69	0.01	0.06	0.55	0.05	2.57	0.12	0.29	0.15	0.08	93.31	1.65
*	6.93	0.50	1.58	14.93	1.11	1.12	15.26	3.09	3.34	8.69	1.97	7.65	0.63	6.90	2.87	1.23	22.20	0.00

Figure 2: Confusion matrix for L2 Korean vowel pronunciation of Japanese speakers

to L2 error patterns. * represents insertion and deletion. Considering each value as the percentage of observations of the canonical phone row, each row sums to 100.

2.2.3. *Pronunciation error patterns*

Pronunciation error patterns for each L1 background are inferred from the corresponding confusion matrix. Error patterns include substitution, insertion, and deletion errors. We propose two types of analyses to scrutinize L1-influenced error patterns.

In the first analysis, we focus on determining the common error patterns shared across five non-native groups. These common patterns can reveal the general characteristics of L2 Korean speech and be used for core feedback in pronunciation training. The first step involves choosing phones for each L1 background that have an accuracy lower than the average accuracy. Different thresholds are used because the performance of the phone recognizer can vary depending on the L1 background. Then, a list of pronunciation error patterns is created for each L1 background, which includes the three most common error patterns for each selected phone. Lastly, the common error patterns are identified by comparing the five error pattern lists and determining which error patterns are shared.

As for the second analysis, we aim to find unique patterns for each L1. Language-dependent error patterns can give guidelines for providing learners with customized feedback. The Chi-square test of independence is conducted to compare each L1 group to the average of five L1 groups to look at the association of L1 with each error pattern. The four most frequently observed patterns for each canonical phone are only tested, leading to a total of 840 error patterns (5 L1 backgrounds × 42 canonical phones (40 phones, insertion, deletion) × 4 realized phones) being statistically analyzed.

For both analyses, native confusion matrix error values were subtracted from corresponding non-native confusion matrix values beforehand, as they include not only the errors of the phone recognizer itself but also the pronunciation variation of native speakers.

3. RESULTS

3.1. Common error patterns

Three types of error patterns are observed to be commonly shared across five languages: two types for consonants and one type for vowels. For consonants, seven error patterns are related to the first type, substitution errors of tense or aspirated with plain consonants: /tʰ/to/t/, /kʰ/to/k/, /tɕʰ/to/tɕ/; /p͈/to/p/, /t͈/to/t/, /t͈ɕ/to/tɕ/, /s͈/to/s/. The second error type is deletion errors of syllable-final, which includes [t̚] and /l/ coda deletions. As for vowels, substitution errors from diphthongs to monophthongs are found: /wi/to/i/, /ɰi/to/i/, /wʌ/to/ʌ/.

3.2. Language-dependent error patterns

According to the statistical results, only 39 out of 840 pronunciation patterns showed L1 association. 25 patterns had L1 causing speakers with higher mispronunciation rates than the average, and the rest 14 had L1 being helpful to better pronunciations. In this paper, we focus on the former 25 pronunciation error patterns as presented in Table 2 and show the full results in our repository[1]. *Canon* and *Real* each refers to canonical and realized phones, with *Average Freq* referring to the confusion matrix value averaged across the five L1s. For example, row 1 can be interpreted that the [k̚] accuracy of Chinese learners was lower than the average of five L1s.

Chinese-speaking learners have 3 language-dependent error patterns including [k̚] mispronunciation, [k̚] deletion, and /ju/to/jo/ substitution. **Vietnamese**-speaking learners have 3 error patterns which includes substitution from /t͈ɕ/to/tɕʰ/, /l/ mispronunciation, and the following /l/to/n/ substitution. **Japanese**-speaking learners have the most various association with L1, with 18 errors showing statistical significance. The error patterns include mispronounced aspirated stops (/p^h/,/t^h/,/k^h/) and their substitution to tense stops (/p͈/,/t͈/,/k͈/). Mispronunciation of /ŋ/ and its substitution to /n/ are more frequent than average learners. Insertion of /ɾ/ and /ɯ/ and substitution of diphthongs to monophthongs are also observed to be significantly frequent (/wɐ/,/wɛ/,/wi/). Lastly, the learners often mispronounce monophthongs (/u/,/ʌ/) to different monophthongs (/ɯ/,/o/). For **Thai**-speaking learners, frequent substitution of /tɕ^h/to/s/ are found. No phones are significantly mispronounced by **English**-speaking learners only. This result may imply that English-speaking learners have advantages in learning Korean compared to other learners. However, to validate this claim, further research using a more balanced dataset is essential, considering the limited number of English-speaking participants in our dataset.

Table 2: Frequency values (%) of confusion matrices. * < .05, ** < .01, *** < .001

L1	Canon	Real	L1 Freq. (%)	Average Freq. (%)	P-value
ZH	[k̚]	[k̚]	58.87	**74.95**	*
	[k̚]	*	**22.79**	7.32	**
	/ju/	/jo/	**23.50**	11.61	*
VI	/tɕ͈/	/tɕ^h/	**18.73**	7.99	*
	/l/	/l/	56.54	**73.23**	*
	/l/	/n/	**23.20**	7.52	**
JP	/p^h/	/p^h/	51.22	**70.70**	**
	/p^h/	/p͈/	**20.72**	7.33	*
	/t^h/	/t^h/	44.40	**62.24**	*
	/t^h/	/t͈/	**20.51**	5.99	**
	/k^h/	/k^h/	54.21	**70.20**	*
	/k^h/	/k͈/	**31.15**	12.74	**
	/ŋ/	/ŋ/	55.22	**82.67**	***
	/ŋ/	/n/	**31.22**	10.10	***
	*	[ɾ]	**10.43**	1.17	*
	*	/ɯ/	**8.98**	0.07	**
	/wɐ/	/ɐ/	**14.58**	3.96	*
	/wɛ/	/ɛ/	**26.69**	9.42	**
	/wi/	/wi/	43.28	**58.58**	*
	/wi/	/i/	**32.05**	11.49	***
	/u/	/u/	64.03	**77.79**	*
	/u/	/ɯ/	**24.72**	7.08	**
	/ʌ/	/ʌ/	66.99	**81.78**	*
	/ʌ/	/o/	**25.58**	10.46	**
TA	/tɕ^h/	/s/	**22.05**	7.33	**

4. DISCUSSION

Based on the observations in Section 3, we uncover the three L1-influence factors attributing to the L2 Korean pronunciation error patterns. The result of our analysis is found to be further supported by the conclusions from various previous literature.

First is differences in **phoneme inventory** [18, 19]. Substitution errors of three-way contrasted consonants can be explained by languages not distinguishing consonants as plain-tense-aspirated [20]. The commonly found substitutions to plain consonants further support the analysis, regarding each five languages includes plain consonants. Results also reveal an interesting finding of Japanese learners' consistent tensification of aspirated stops, alongside the tendency to substitute aspirated stops to plain. Previously, word-medial tense and aspirate distinction in perception have been reported to be difficult for Japanese speakers [21]. Our study implies that the production of aspirated stops may also be related to this phenomenon.

Error patterns can also be attributed to differences in **syllable structure**, especially in the syllable-final position of languages [22]. While Korean allows /l/ as coda, four languages excluding English do not allow syllable-final /l/. Although English allows /l/ as the final consonant, the allophone /ɫ/ at the word-final position sounds different from Korean /l/. Vietnamese learners especially had a higher chance of mispronouncing /l/, in particular as /n/ that has a similar place of articulation [23, 24]. Chinese learners show a significant pattern of [k̚] deletion, which was previously explained in [25] with syllable structure. Japanese learners consistently insert /ɯ/ as well, due to their syllable-based constraints with the tendency to open-syllabification [26].

Lastly, **pronunciation rule** can explain the frequent deletion of coda [t̚]. The corresponding rule is tensification, where the plain consonants are pronounced as tense in a stop-plain consonant sequence. While the sequence must be acoustically realized as long closure, many L2 Korean learners are reported to produce shorter closures compared to the natives, because they are either unaware of or unfamiliar with the tensification rule [27, 28]. The mispronounced sounds are often perceived as deletion of coda [27], which is observed in our automatic phone transcriptions as well.

5. CONCLUSION

This paper presents a large-scale linguistic analysis on L2 Korean pronunciation error patterns from speakers of five different L1s, Chinese, Vietnamese, Japanese, Thai, and English, by using the automatic phonetic transcription. Comparative analyses regarding L1 background reveal common error patterns and L1-dependent error patterns, which can be further explained by three types of L1 influences: phoneme inventory, syllable structure, and pronunciation rule. This study contributes not only in integrating and expanding the previous knowledge of L2 Korean error patterns on a large-scale basis, but also validating the possibilities of automatic phonetic transcriptions on non-native speech analysis. Future work includes considering L1 and uncategorizable phones to yield a better and deeper understanding of L1-influenced non-native pronunciation error patterns.

ACKNOWLEDGMENT

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-2018-0-01833) supervised by the IITP (Institute for Information&Communications Technology Planning&Evaluation)

6. REFERENCES

[1] P. M. Rogerson-Revell, "Computer-assisted pronunciation training (capt): Current issues and future directions," *RELC Journal*, vol. 52, no. 1, pp. 189–205, 2021.

[2] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, "3m: An effective multi-view, multi-granularity, and multi-aspect modeling approach to english pronunciation assessment," in *2022 APSIPA ASC*, 2022, pp. 575–582.

[3] M. Yang, K. Hirschi, S. D. Looney, O. Kang, and J. H. L. Hansen, "Improving mispronunciation detection with wav2vec2-based momentum pseudo-labeling for accentedness and intelligibility assessment," in *Interspeech*, 2022.

[4] L. Peng, K. Fu, B. Lin, D. Ke, and J. Zhang, "A study on fine-tuning wav2vec2.0 model for the task of mispronunciation detection and diagnosis," in *Interspeech*, 2021.

[5] E. J. Kong, S. Kang, and M. Seo, "The acoustic cue-weighting and the l2 production-perception link: A case of english-speaking adults' learning of korean stops," *Phonetics and Speech Sciences*, vol. 14, no. 3, pp. 1–9, 2022.

[6] S. Benjamin and H.-Y. Lee, "The perception and production of korean vowels by egyptian learners," *Phonetics and Speech Sciences*, vol. 13, no. 4, pp. 23–34, 2021.

[7] A. Jatteau, I. Vasilescu, L. Lamel, and M. Adda-Decker, "Final devoicing of fricatives in french: Studying variation in large-scale corporawith automatic alignment," 2019.

[8] J. Yang and R. Fox, "L1-l2 interactions of vowel systems in young bilingual mandarin-english children," *J. Phonetics*, vol. 65, pp. 60–76, 2017.

[9] J. Yuan, H. Lin, and Y. Liu, "Nasal coarticulation in l1 and l2 english speech: A large-scale study," *Training*, pp. 96–6, 1974.

[10] S. Shi and C. Shih, "Acoustic analysis of l1 influence on l2 pronunciation errors: A case study of accented english speech by chinese learners," *ICPHS, Melbourne, Australia*, 2019.

[11] S. Shi, C. Shih, and J. Zhang, "Capturing l1 influence on l2 pronunciation by simulating perceptual space using acoustic features," in *Interspeech*, 2019, pp. 2648–2652.

[12] C. Cucchiarini and H. Strik, "Automatic phonetic transcription: An overview," in *ICPHS*, 2003, pp. 347–350.

[13] A. Babu *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Interspeech*, 2022.

[14] K. Choi and H.-M. Park, "Distilling a pretrained language model to a multilingual asr model," in *Interspeech*, 2022.

[15] "Foreign speakers' voice data of korean for artificial intelligence learning," https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=505, 2020.

[16] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[17] D. Povey *et al.*, "The kaldi speech recognition toolkit," in *ASRU*, 2011.

[18] L. Wang, X. Feng, and H. Meng, "Mispronunciation detection based on cross-language phonological comparisons," in *ICALIP*, 2008, pp. 307–311.

[19] S. Khanal, M. T. Johnson, and N. Bozorg, "Articulatory comparison of l1 and l2 speech for mispronunciation diagnosis," in *SLT*, 2021, pp. 693–697.

[20] J. J. Holliday, "A longitudinal study of the second language acquisition of a three-way stop contrast," *Journal of Phonetics*, vol. 50, pp. 1–14, 2015.

[21] T. Yasuta, *Stop perception in second language phonology: Perception of English and Korean stops by Japanese speakers*. University of Hawaii, 2004.

[22] Y. He, "Production of english syllable final/l/by mandarin chinese speakers." *Journal of Language Teaching & Research*, vol. 5, no. 4, 2014.

[23] K. B. Choe, J. Y. Song, and S.-G. Park, "Analysis of pronunciation errors of korean syllable-final liquid /ㄹ/ by vietnamese learners of korean," *Urimal*, no. 64, pp. 181–209, 2021.

[24] M.-k. Hwang, "Analysis of the Korean liquid pronunciation of Vietnamese entry-level Students," *Journal of Ehwa Korean Language and Literature*, vol. 39, pp. 49–87, 2016.

[25] E. J. Lee and I. H. Woo, "A study on an educational plan for the pronunciation of the final consonants /ㄱ, ㄷ, ㅂ/ in korean for chinese korean learners -by applying the syllable structure of korean to that of chinese reversely-," *The Academy for Korean Language Education*, no. 97, pp. 327–359, 2013.

[26] E. K. Yoon, "An experimental perceptual study of different syllable structures in l1 & l2," *Journal of Dong-ak Language and Literature*, no. 64, pp. 409–438, 2015.

[27] L. J and P. K, "A study on the korean learners' realization aspect and teaching methods of glottalization after obstruent," *The Korean society of bilingualism*, no. 71, pp. 223–248, 2018.

[28] K. K H *et al.*, "A study on the final consonant pronunciation errors of chinese korean learners," Ph.D. dissertation, Jeju University, 2019.

[1] https://github.com/eunjung31/L2K-PronunciationError