pISSN 2005-8063
eISSN 2586-5854
2026. 3. 31.
Vol.18 No.1
pp. 55-63

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2026.18.1.055

Whisper automatic speech recognition errors in Korean complex coda recognition: Syllable coda phonotactics and dialectal variation

Tae-Jin Yoon[1,‡], Soohyun Kwon[2,‡], Jeong-Im Han[3,*]

[1]*Department of English Language and Literature, Sungshin Women's University, Seoul, Korea*
[2] *School of Global Communication, Kyunghee University, Yongin, Korea*
[3]*Department of English Language and Literature, Konkuk University, Seoul, Korea*

Abstract

This study investigates how Whisper automatic speech recognition models handle Korean syllable-final consonant clusters, focusing on the effects of dialectal variation, model size and phonological variation. We analyzed 54,536 spontaneous-speech utterances from the National Institute of Korean Language (NIKL) Daily Conversation Corpus 2022, comparing speakers from Seoul and Gyeongsang (consolidating Busan, Daegu, Gyeongnam, Gyeongbuk, and Ulsan) across four Whisper variants (small, medium, large-v2, large-v3). Overall coda error rates ranged from 2.67% (large-v3) to 4.53% (small), while consonant cluster error rates were substantially higher, ranging from 2.94% (large-v3, Seoul) to 6.93% (small, Gyeongsang). Seoul consistently showed lower error rates than Gyeongsang, but this advantage narrowed with model size and disappeared in large-v3. Phonologically-motivated errors constituted a consistent and substantial minority of all errors (23%–41% across models and varieties), with a statistically robust subtype hierarchy: C2 simplification was dominant, followed by C1 simplification and aspiration merger, while nasal assimilation and resyllabification occurred at near-zero rates. A Seoul-over-Gyeongsang asymmetry in C1 simplification errors, significant in smaller models and attenuating with scale, is attributed to an interaction between Whisper's Seoul-dominant training data and ongoing C1-retention change. These findings demonstrate that automatic speech recognition (ASR) errors in morphophonologically complex languages are not random but represent systemmatic failure points rooted in phonological grammar, and that incorporating explicit phonological knowledge into training or fine-tuning pipelines may offer a more targeted and linguistically motivated strategy for improving consonant cluster recognition than simply scaling model size.

Keywords: automatic speech recognition (ASR), Whisper, Korean complex codas (orthographic double codas), phonological structure, prosodic structure, dialectal variation, coda simplification

‡These authors contributed equally and share first authorship.

* jhan@konkuk.ac.kr, Corresponding author

Received 13 February 2026; Revised 12 March 2026; Accepted 14 March 2026

1. Introduction

Recent end-to-end automatic speech recognition (ASR) systems commonly adopt transformer-based sequence-to-sequence architectures that predict a sequence of text tokens (i.e., orthographies) directly from acoustic input. Whisper (Radford et al., 2023) is trained on a large-scale multilingual speech–text data and produces transcriptions by combining segmental or prosodic evidence from the acoustic signal with contextual (language-model) information during decoding.

Although the Korean orthographic system ('Hangul') is relatively transparent in that spelling-to-sound correspondence is consistent, Korean poses a persistent challenge for such systems owing to the tautosyllabic consonant clusters in the syllable-coda position. Given that Korean does not allow consonant clusters in the coda as well as onset positions, underlying consonant clusters undergo various morphophonological processes. For example, when they occur in isolation forms or before another consonant across the morpheme boundary, they undergo consonant cluster simplification (CCS) and coda neutralization (e.g., /talk/ 'chicken' + /to/ 'too' > [tak$^{\neg}$.to]). When they are followed by a vowel across the morpheme boundary, the second member of the cluster is resyllabified onto the onset of the following syllable (e.g., /talk/ 'chicken' + /i/ nominative marker > [tal.ki]) (Jun, 1998; NIKL, 2017; Sohn, 1999).

Moreover, Korean CCS patterns are conditioned by cluster type, dialect, speaker age and lexical items. Variation is observed across cluster types, with some clusters preserving C1 and others preserving C2 (e.g., /tols/→[tol], /salm/→[sam]). Early studies have noted the differences in the realization of *l*-initial clusters depending on dialects such that Seoul Korean tends to retain C2 while Gyeongsang dialects tend to retain C1 (Cho, 1988; Whitman, 1985). Interestingly, later experimental works have shown that for /lk/ and /lp/, preserving [l] is becoming increasingly prevalent (Cho & Kim, 2009; Kim & Kang, 2021). More recently, Kwon et al. (2023, 2025), based on a large-scale spontaneous speech corpus, have proposed that the observed patterns for /lk, lp, lm/ represent intermediate stages of an ongoing analogical change, with considerable variation across lexical items.

From an ASR perspective, morphophonological alternations and complex patterns of variation make Korean consonant clusters particularly susceptible to recognition errors. In addition to categorical morphophonological alternations, Korean coda consonants are typically unreleased and can trigger phonetic alternations such as post-obstruent tensing or place/voicing assimilation in the following consonant, which further obscures segmental cues for cluster members. Because Whisper optimizes orthographic rather than phonological reconstruction, our evaluation focuses on categorical mappings among orthographic coda categories, while recognizing that gradient phonetic variation can underlie some confusions.

2. Theoretical Background

2.1. Alterations Involved in the Realization of Consonant Clusters in Korean.

There are several morphophonological processes that may potentially be involved in recognizing Korean consonant clusters. In this section, we review these processes to assess the extent to which phonologically motivated errors account for the errors produced by Whisper.

Because Korean does not permit consonant clusters in the syllable coda on the surface, stem-final clusters (C1C2) are typically reduced to a single consonant when followed by consonant-initial morphemes. Previous findings have discovered that one of the two consonants (C1 or C2) survives, which is conditioned by cluster type, dialect region, gender, and age. For example, stop–initial clusters such as /ps/ and /ks/ and nasal-initial clusters such as /nc/ and /nh/ are invariably realized as C1, whereas some liquid-initial clusters such as /lm/ and /lp^{h}/ favor C2 retention in certain varieties. In contrast, in /lp/ and /lk/, younger speakers are likely to preserve C1 or even both consonants (C1C2) of the cluster, although these patterns depend on dialects and lexical items (Cho & Kim, 2009; Kim & Kang, 2021, Kwon et al., 2025).

When a vowel-initial morpheme follows, the clusters undergo resyllabification: C2 is licensed as the onset of the following syllable, yielding [C1.C2V] on the surface. In that case, the surface form can show liaison-like sequencing (e.g., [C1.C2V]), which can lead to the preservation of two consonantal gestures on the surface (Jun, 1998; Kim, 2022).

Yet another alternation involved is an aspiration merger whereby a lax stop becomes aspirated before /h/ (progressive aspiration merger) as in /palk-hi-ko/ 'being lit and'→[pal.k^{h}i.ko] or an underlying /h/ triggers aspiration when it is followed by a stop consonant (regressive aspiration merger) as in /ilh-ko/ 'lose and'→[il.k^{h}o] (e.g. Kim-Renaud, 1974).

Also, the process of obstruent nasalization is relevant in that once the cluster is simplified to C2, an obstruent, and this syllable-final obstruent in Korean becomes a nasal when the following syllable begins with a nasal (Cho, 2016; Kim-Renaud, 1974; Sohn, 1999). For example, in /ilk-nin/→ ik.nin→[iŋ.nin] 'reading', an underlying C_2 /k/ is nasalized on the surface.

2.2. Dialectal Variation

The patterns of cluster simplification have long been reported to vary by dialects (Cho, 1988; Sohn, 1999; Whitman, 1985). For /lk/ and /lp/, the Seoul dialect has been described as preferring C2 retention ([k], [p]), whereas Gyeongsang varieties have been described as preferring C1 retention ([l]) (Cho, 1988; Whitman, 1985). However, more recent work suggests that this system is not static: younger speakers show increased C1 retention and/or two-consonant realizations for liquid-initial clusters, and such a change can proceed via lexical diffusion and later generalization by analogy (Cho & Kim, 2009; Kim & Kang, 2021; Kwon et al., 2025).

3. The Present Study

This study investigates the rates and patterns of errors that Whisper exhibits when recognizing Korean syllable-final consonant clusters and examines how error patterns vary as a function of dialect, model size, and morphophonological process. We use a large-scale spontaneous-speech dataset, the NIKL Daily Conversation Corpus 2022 (Version 1.0), sampling speakers from

Seoul and Gyeongsang. Orthographic reference transcriptions are compared with the outputs produced by four Whisper model variants.

The specific research questions to be addressed are as follows:

RQ1: What is the overall error rate for recognizing Korean consonant clusters in Whisper as compared to simple coda, and how does it vary depending on dialect and model size?

RQ2: What types of recognition errors occur in Korean consonant clusters, how frequently do they occur, and what systematic patterns emerge in their distribution?

Based on the previous findings, we make the following predictions:

Prediction 1 (dialect and model size): Error rates involving consonant clusters are expected to be higher for Gyeongsang dialects than for the Seoul dialect, reflecting dialectal divergence from the Seoul-centered orthographic standard and the training data. Specifically, we predict that Gyeongsang speakers' tendency to retain C1, compared to the Seoul speakers' preference for C2 in consonant clusters such as /lk/ and /lp/, will result in greater mismatches with Seoul-based orthographic transcriptions, leading to higher error rates in Gyeongsang.

Prediction 2 (error types): Recognition errors will be modulated by morphophonological variations in Korean for the l-initial clusters with variable realizations (e.g., /lk/, /lm/, /lp/) and the resulting consonant cluster simplification phenomenon. Error types are expected to align with phonologically motivated alternations.[2]

To assess these predictions, we compare Whisper outputs to reference transcripts. Although an ASR output matches the spoken surface form faithfully (e.g., innovative simplification in prevocalic position; /talk + i/ > [ta.ki]), it will be counted as an "error" if the reference maintains the orthographic consonant cluster.

4. Methods

4.1. Data and Preprocessing

We use the NIKL Daily Conversation Speech Corpus 2022 (v1.0). This corpus consists of 16 topics of everyday conversations and 10 topics of cooperative dialogues. Each session involves 2–4 speakers interacting for roughly 15 minutes; the corpus documentation reports approximately 2,000 speakers and about 630 hours of recordings. Transcriptions are organized in intonation-phrase units and provide both an orthographic tier (standard spelling) and a pronunciation tier (speech-based transcription). The release date for the 2022 corpus (v1.0) is 29 December 2023.

We first assigned speakers to two regional groups based on speaker metadata (birthplace/residence): (1) Seoul (n=27,620 utterances), and (2) Gyeongsang (n=26,916 utterances), consolidating five Gyeongsang-speaking (Busan, Daegu, Gyeongnam, Gyeongbuk, Ulsan). We then filtered intonation-phrase utterances to retain only those containing at least one "eojeol" with a syllable-coda cluster. The final dataset contains 54,536 intonation-phrase utterances. As the audio materials are distributed as PCM files, we converted them to 16 kHz, single-channel WAV files to match Whisper's input requirements.

Table 1. Distribution of regional dialect data

	Seoul	Gyeongsang	Total
Audio files	6,905	6,729	13,634
Utterances	27,620	26,916	54,536
Coda syllables	365,870	345,471	711,341
Percentage of coda syllables (%)	51.4	48.6	100

Gyeongsang includes Busan (3,280 audio files), Daegu (1,616), Gyeongnam (873), Gyeongbuk (676), and Ulsan (284).

As in Table 1, the data was collected from spontaneous speeches (e.g., free conversation, interviews, task-oriented dialogues) and thus include naturally occurring phonological variations and dialectal characteristics.

4.2. Whisper Models and Decoding Settings

We evaluated four Whisper model variants: (1) whisper-small (244M parameters), (2) whisper-medium (769M), (3) whisper-large-v2 (1.55B), and (4) whisper-large-v3 (1.55B). These checkpoints span the major publicly released Whisper model sizes, and the inclusion of both large-v2 and large-v3 (same parameter count but different training/version updates) allows us to test whether version changes affect Korean consonant cluster recognition. All models were run in Korean transcription mode with default decoding parameters. For each utterance, we generated four independent transcriptions, resulting in 218,144 outputs (54,536×4).

4.3. Coda-Level Analysis Pipeline

To analyze coda discrepancies between the ASR outputs and the reference transcriptions, we decomposed each Hangul syllable into onset, nucleus, and coda components. The target coda is represented orthographically as one of 28 categories (ø, ㄱ, ㄲ, ㄳ, ㄴ, ㄵ, ㄶ, ㄷ, ㄹ, ㄺ, ㄻ, ㄼ, ㄽ, ㄾ, ㄿ, ㅀ, ㅁ, ㅂ, ㅄ, ㅅ, ㅆ, ㅇ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ), including "no coda" (ø); a subset corresponds to orthographic consonant clusters (ㄳ, ㄵ, ㄶ, ㄺ, ㄻ, ㄼ, ㄽ, ㄾ, ㅀ, ㅄ). We aligned the reference and ASR syllable sequences using Python's difflib.SequenceMatcher to obtain syllable-level correspondences.

The primary objective of Whisper is not to derive the surface phonetic representation but to recover the intended spelling. Accordingly, in this study, we define ASR errors as discrepancies between the ASR outputs and the orthographic transcript. For example, the word /alh-ta/ 앓다 'suffer from (a disease)' may be phonetically realized as [alᵗa] and recognized as such by Whisper, but the result fails to recover the intended spelling. We therefore treat such cases as recognition errors in

2 The notation '< >' was used to clearly indicate spellings, while phonemes and phonetic forms were represented by placing the corresponding phonetic symbols within the slant brackets (/ /) and square brackets ([ ]), respectively.

our evaluation.

We classified all tokens containing consonant clusters broadly into three categories: (1) Match: the tokens that are correctly recognized, where the ASR output matches the reference transcription and recovers the intended spelling; (2) Phonologically-motivated errors: the recognition errors that result from the application of morphophonological processes in Korean; (3) Spurious errors: the tokens where mismatch exists but follows no phonological rule (e.g., ASR over-generates a cluster, or two different clusters are swapped).

Phonologically motivated errors are further categorized into five subtypes according to the morphophonemic processes reflected in the output, as discussed in Section 2.1. The first subtype is C1 simplification error, in which the output fails to recover the underlying consonant cluster and retains only C2, as in /salm-to/ 'life too'→[sam.to]. The second subtype is C2 simplification error, in which the output fails to recover the underlying consonant cluster and retains only C1, as in /palp-ta/ 'step on'→[pal.ta]. The third subtype is resyllabification error, in which the output fails to recover the underlying consonant cluster and the second member of the cluster is resyllabified into the onset of the following syllable, as in /ilk-ʌ-sʌ/ 'because of reading'→[il.kʌ.sʌ]. The fourth subtype is aspiration merger error, in which a lax stop becomes aspirated before /h/ (progressive aspiration merger), as in /palk-hi-ko/ 'being lit and' →[pal.khi.ko], or an underlying /h/ triggers aspiration when it is followed by a stop consonant (regressive aspiration merger), as in /ilh-ko/ 'lose and'→[il.kho]. The fifth subtype is obstruent nasalization error, in which a syllable-final obstruent becomes nasalized when the following syllable begins with a nasal consonant. For example, in /ilk-nɨn/ 'reading', the underlying C2 /k/ undergoes nasalization on the surface: /ilk-nɨn/→ik.nɨn→ [iŋ.nɨn].

With these error types defined, coda error rate was computed as follows:

$$\text{Coda error rate} = \text{No. of errors} / (\text{No. of matches} + \text{errors}) \times 100\%$$

5. Results

5.1. Overall Coda Recognition Error Rates

Table 2 presents mean coda (single and complex codas combined) recognition error rates by regional group across the four Whisper model variants. Both Seoul and Gyeongsang varieties show a clear and consistent improvement with increasing model size: error rates decline from 4.10% (Seoul) and 4.74% (Gyeongsang) in the small model to 2.74% and 2.89%, respectively, in the large-v3 model, representing a reduction of approximately one third across the model scale. Chi-square tests of independence confirmed that Seoul outperformed Gyeongsang significantly at every model size [small: $\chi^2(1)=172.52$, $p<.001$, $\varphi=.016$; medium: $\chi^2(1)=20.40$, $p<.001$, $\varphi=.005$; large-v2: $\chi^2(1)=34.48$, $p<.001$, $\varphi=.007$; large-v3: $\chi^2(1)=14.61$, $p<.001$, $\varphi=.005$], though effect sizes were consistently small across all models (all $\varphi<.02$), indicating that while the regional differences are statistically reliable, their practical magnitude is limited.

Table 2. Mean coda recognition error rate (%) by regional group and model size

Region	small	medium	large-v2	large-v3
Seoul	4.1	3.16	2.95	2.74
Gyeongsang	4.74	3.35	3.19	2.89
Difference	+0.54	+0.19	+0.24	+0.15

Pairwise chi-square tests with Bonferroni correction (adjusted $\alpha=.008$ for 6 comparisons) confirmed that every step between model sizes yielded a statistically significant improvement in both varieties (all $ps<.001$). Notably, even the incremental gains within the larger models—medium to large-v2 and medium to large-v3—were reliable in both Seoul and Gyeongsang, indicating that scaling benefits persist throughout the full model range. The regional gap, however, does not decrease monotonically: it is widest in the small model (+0.54%p), narrows sharply to +0.19%p at medium, widens slightly at large-v2 (+0.24%p), and reaches its minimum at large-v3 (+0.15%p). This non-monotonic trajectory suggests that the relationship between model capacity and dialect robustness is not strictly linear, though the overall trend clearly favors convergence at larger model sizes.

5.2. Regional Differences in Consonant Cluster Recognition Error Rates

Table 3 presents mean consonant cluster recognition error rates by regional group across the four Whisper model sizes. A clear improvement is observed with increasing model capacity in both varieties: error rates decline from 6.07% (Seoul) and 6.93% (Gyeongsang) in the small model to 2.94% and 3.15%, respectively, in the large-v3 model—representing an approximate 50% relative reduction across the model scale.

Table 3. Mean consonant cluster recognition error rate (%) (as compared to simple coda recognition error rate) by regional group and model size

Region	small	medium	large-v2	large-v3
Seoul	6.07	3.71	3.14	2.94
Gyeongsang	6.93	4.12	4.06	3.15
Difference	+0.86	+0.41	+0.92	+0.21

Chi-square tests of independence revealed that the Seoul–Gyeongsang difference was statistically significant for the small model [$\chi^2(1)=4.13$, $p=.042$, $\varphi=.017$] and the large-v2 model [$\chi^2(1)=8.22$, $p=.004$, $\varphi=.025$], but did not reach significance for the medium model [$\chi^2(1)=1.52$, $p=.218$] or the large-v3 model [$\chi^2(1)=0.51$, $p=.474$]. Effect sizes were consistently small across all models (all $\varphi<.03$), indicating that while regional differences are statistically detectable at certain model sizes, their practical magnitude is limited. Notably, the large-v2 model showed the largest regional gap (0.92%p) and the strongest effect, whereas the large-v3 model showed the smallest gap (0.21%p) and no significant difference, suggesting that the largest model effectively closes the dialectal performance gap in consonant cluster recognition.

Pairwise chi-square tests with Bonferroni correction (adjusted $\alpha=.008$ for 6 comparisons) were conducted to assess whether improvements across model sizes were statistically reliable within each variety. In both Seoul and Gyeongsang, all comparisons

involving the small model were highly significant (all *p*s<.001), confirming that the step from small to any larger model yields a meaningful reduction in error rate. However, the two varieties diverged in their patterns of improvement beyond the small model. In Seoul, neither the medium-to-large-v2 nor the large-v2-to-large-v3 comparison reached significance (*p*=.068 and *p*=.488, respectively), and medium-to-large-v3 only approached but did not survive correction (*p*=.012), suggesting that error rate gains in Seoul effectively plateau after the medium model. In Gyeongsang, by contrast, both medium-to-large-v3 (*p*=.003) and large-v2-to-large-v3 (*p*=.005) were significant after Bonferroni correction, indicating that continued scaling yields further reliable gains for the Gyeongsang variety even at larger model sizes. Effect sizes remained small throughout (all φ<.09), but this asymmetric scaling pattern suggests that Gyeongsang speech benefits more from the full range of model capacity than Seoul speech does.

5.3. Recognition Error Patterns of Consonant Clusters

We examined the patterns of consonant cluster recognition errors in detail for the Seoul and Gyeongsang dialects under each model.

Table 4. Distribution of ASR error types by region—Whisper small model

Error type	Subtype	Seoul		Gyeongsang	
		N	%	N	%
Phonologically motivated errors	C2 Simplification	90	21.5	97	20.8
	C1 Simplification	37	8.8	30	6.4
	Aspiration merger	27	6.4	23	4.9
	Nasal assimilation	5	1.2	2	0.4
	Resyllabification	0	0	2	0.4
Phonologically motivated errors		159	37.9	154	33
Spurious errors		260	62.1	312	67
Total		419	100	466	100

Table 4 presents the distribution of ASR error types across the two regional varieties, Seoul and Gyeongsang under the small model. For both varieties, spurious errors constituted the majority of all errors, accounting for 65.6% of the Seoul total (N=718) and 69.0% of the Gyeongsang total (N=847). Still, phonologically motivated errors made up the remaining 34.4% (N=377) and 31.0% (N=381) for Seoul and Gyeongsang, respectively, suggesting that the model's errors reflect underlying phonological processes to a non-trivial degree in both varieties. The difference in the overall proportion of phonologically motivated versus spurious errors between the two varieties did not reach statistical significance [$\chi^2(1)=3.05$, $p=.081$].

Among the phonologically motivated subtypes, C2 simplification was by far the most frequent in both groups, comprising 19.3% of all Seoul errors (N=211) and 18.7% of all Gyeongsang errors (N=239). C1 simplification was the second most common subtype overall, though it was notably more frequent in Seoul (8.3%, N=91) than in Gyeongsang (5.2%, N=66). Aspiration merger occurred at similar rates in both varieties, yielding 6.1% in Seoul and 5.2% in Gyeongsang. Nasal assimilation was rare in both groups (Seoul: 0.7%, N=8; Gyeongsang: 0.5%, N=6). Resyllabification errors were absent in the Seoul data entirely, while a small number were attested in GS (0.3%, N=4).

A chi-square goodness-of-fit test confirmed that the distribution of phonologically motivated errors across subtypes was highly unequal in both varieties [Seoul: $\chi^2(3)=231.54$, $p<.001$; Gyeongsang: $\chi^2(4)=486.77$, $p<.001$], reflecting a strong concentration of errors in C2 simplification. Pairwise Fisher's exact tests with Bonferroni correction (adjusted α=.005 for 10 comparisons) revealed a robust and consistent subtype hierarchy across both varieties. C2 simplification was significantly more frequent than every other subtype in both Seoul and Gyeongsang (all *p*s<.001). C1 simplification and aspiration merger did not differ significantly from each other in Gyeongsang (*p*=.923), though in Seoul C1 simplification occurred at a marginally higher rate than aspiration merger (*p*=.039), a difference that did not survive Bonferroni correction. Both subtypes were significantly more frequent than nasal assimilation (all Bonferroni-corrected *p*s<.001) and resyllabification (all *p*s<.001). The contrast between nasal assimilation and resyllabification was significant in Seoul (*p*=.008) but not in Gyeongsang (*p*=.752), where both subtypes were uniformly rare or absent.

Between-variety comparisons for individual subtypes revealed one significant difference: C1 simplification occurred at a significantly higher rate in Seoul than in Gyeongsang (*p*=.006), while all other subtype comparisons were non-significant (C2 simplification: *p*=.916; aspiration merger: *p*=.368; nasal assimilation: *p*=.593; resyllabification: *p*=.127).

Taken together, the distributions suggest that while the overall proportions of phonologically motivated versus spurious errors are broadly comparable across the two varieties, Seoul exhibits a slightly higher rate of phonologically motivated errors overall, driven in particular by a significantly higher incidence of C1 simplification.

Table 5. Distribution of ASR error types by region—Whisper medium model

Error type	Subtype	Seoul		Gyeongsang	
		N	%	N	%
Phonologically motivated errors	C2 Simplification	43	16.8	52	18.8
	C1 Simplification	21	8.2	12	4.3
	Aspiration merger	18	7	19	6.9
	Nasal assimilation	2	0.8	2	0.7
	Resyllabification	0	0	2	0.7
Phonologically-motivated errors		84	32.8	87	31.4
Spurious errors		172	67.2	190	68.6
Total		256	100	277	100

Table 5 presents the distribution of ASR error types produced by the medium model across Seoul and Gyeongsang. Of the 256 Seoul tokens and 277 Gyeongsang tokens in the dataset, spurious errors constituted the majority in both varieties, accounting for 67.2% (N=172) and 68.6% (N=190), respectively. Still, phonologically motivated errors made up the remaining 32.8% (N=84) for Seoul and 31.4% (N=87) for Gyeongsang, indicating

broadly comparable overall error profiles across the two groups. This difference did not reach statistical significance [$\chi^2(1)=0.12$, $p=.729$], confirming that the medium model shows no reliable regional asymmetry at the level of broad error type.

Among the phonologically motivated subtypes, C2 simplification was the most frequent category in both varieties, representing 16.8% of Seoul errors (N=43) and 18.8% of GS errors (N=52). C1 simplification ranked second in both groups, though it occurred at a notably higher rate in Seoul (8.2%, N=21) than in Gyeongsang (4.3%, N=12), suggesting a dialectal asymmetry. Aspiration merger was comparable across varieties (Seoul: N=18, 7.0%; Gyeongsang: N=19, 6.9%). Nasal assimilation was rare in both groups (Seoul: 0.8%, N=2; Gyeongsang: 0.7%, N=2). Resyllabification errors were entirely absent in the Seoul data, while two instances were attested in Gyeongsang (0.7%).

A chi-square goodness-of-fit test confirmed that the subtype distribution within phonologically motivated errors was significantly non-uniform in both varieties [Seoul: $\chi^2(3)=40.67$, $p<.001$; Gyeongsang: $\chi^2(4)=97.89$, $p<.001$], with errors concentrated predominantly in C2 simplification. Pairwise Fisher's exact tests with Bonferroni correction (adjusted $\alpha=.005$ for 10 comparisons) revealed a consistent subtype hierarchy across both varieties. C2 simplification was significantly more frequent than all other subtypes in both Seoul and Gyeongsang (all ps<.001). C1 simplification and aspiration merger did not differ significantly from each other in either variety (Seoul: $p=.715$; Gyeongsang: $p=.234$), forming a non-differentiated intermediate tier in both groups. In Seoul, both subtypes were significantly more frequent than nasal assimilation (all ps<.001) and resyllabification (all ps<.001). In Gyeongsang, C1 simplification exceeded nasal assimilation and resyllabification at the nominal level ($p=.010$ for both) but did not survive Bonferroni correction; aspiration merger, however, did significantly exceed both (all ps<.001, Bonferroni-corrected). Nasal assimilation and resyllabification did not differ from each other in either variety (Seoul: $p=.497$; Gyeongsang: $p=1.000$).

Between-variety comparisons for individual subtypes revealed no statistically significant regional differences. The observed Seoul advantage in C1 simplification approached but did not reach significance ($p=.073$), and all remaining subtypes were non-significant (C2 simplification: $p=.573$; aspiration merger: $p=1.000$; nasal assimilation: $p=1.000$; resyllabification: $p=.500$).

Overall, the medium model's error distribution mirrors the general pattern observed across model sizes, with C2 simplification consistently dominating the phonologically motivated subtype category. The directional Seoul–Gyeongsang asymmetry in C1 simplification, while consistent with findings from the small model, does not reach significance under the medium model, suggesting that this effect may be attenuated at larger model sizes.

Table 6. Distribution of ASR error types by region—Whisper large-v2 model

Error type	Subtype	Seoul		Gyeongsang	
		N	%	N	%
Phonologically motivated errors	C2 Simplification	39	18	57	20.9
	C1 Simplification	15	6.9	14	5.1
	Aspiration merger	13	6	18	6.6
	Nasal assimilation	1	0.5	2	0.7
	Resyllabification	0	0	0	0
Phonologicallymotivated errors		68	31.3	91	33.3
Spurious errors		149	68.7	182	66.7
total		217	100	273	100

Table 6 presents the distribution of ASR error types produced by the Whisper large-v2 model across the Seoul and Gyeongsang varieties. A chi-square goodness-of-fit test confirmed that the distribution of phonologically motivated errors across subtypes was significantly unequal in both varieties [Seoul: $\chi^2(3)=44.71$, $p<.001$; Gyeongsang: $\chi^2(3)=74.85$, $p<.001$], indicating that certain subtypes occurred at substantially higher rates than others.

Pairwise Fisher's exact tests with Bonferroni correction (adjusted $\alpha=.005$ for 10 comparisons) revealed a consistent hierarchy of subtype frequencies. C2 simplification was the dominant subtype in both varieties (Seoul: 18.0%, N=39; Gyeongsang: 20.9%, N=57), occurring significantly more frequently than all other subtypes in both Seoul and Gyeongsang (all ps<.001). C1 simplification (Seoul: 6.9%, N=15; Gyeongsang: 5.1%, N=14) and aspiration merger (Seoul: 6.0%, N=13; Gyeongsang: 6.6%, N=18) did not differ significantly from each other in either variety (Seoul: $p=.832$; Gyeongsang: $p=.560$). Both C1 simplification and aspiration merger, however, occurred significantly more frequently than nasal assimilation (all ps≤.003) and resyllabification (all ps<.001). Nasal assimilation (Seoul: 0.5%, N=1; Gyeongsang: 0.7%, N=2) and resyllabification (N=0 in both varieties) did not differ significantly from each other (ps=1.000 and .497 for Seoul and Gyeongsang respectively), as both were at floor.

In sum, the large-v2 model's phonologically motivated error pattern is characterized by a clear and statistically robust dominance of C2 simplification, followed by a non-significantly differentiated middle tier of C1 simplification and aspiration merger, with nasal assimilation and resyllabification occurring at negligible rates.

Table 7. Distribution of ASR error types by region—Whisper large-v3 model

Error type	Subtype	Seoul		Gyeongsang	
		N	%	N	%
Phonologically motivated errors	C2 Simplification	39	19.2	33	15.6
	C1 Simplification	18	8.9	10	4.7
	Aspiration merger	9	4.4	6	2.8
	Nasal assimilation	0	0	0	0
	Resyllabification	0	0	0	0
Phonologically motivated errors		66	32.5	49	23.1
Spurious errors		137	67.5	163	76.9
TOTAL		203	100	212	100

Table 7 presents the distribution of ASR error types produced by the large-v3 model across the Gyeongsang varieties. The total number of tokens was 203 for Seoul and 212 for Gyeongsang. Spurious errors constituted the majority in both varieties (Seoul: 67.5%, N=137; Gyeongsang: 76.9%, N=163). Still, phonologically motivated errors accounted for a large proportion, 32.5% (N=66) of Seoul errors and 23.1% (N=49) of Gyeongsang errors. A chi-square test of independence showed that this difference was statistically significant [$\chi^2(1)=4.57$, $p=.033$], indicating that the large-v3 model produced proportionally more phonologically-motivated errors in Seoul than in Gyeongsang—the most pronounced regional asymmetry observed in the present dataset at the level of broad error type.

A chi-square goodness-of-fit test confirmed that the distribution of phonologically-motivated errors across the three attested subtypes (C2 simplification, C1 simplification, aspiration merger) was significantly unequal in both varieties [Seoul: $\chi^2(2)=21.55$, $p<.001$; Gyeongsang: $\chi^2(2)=26.00$, $p<.001$], with errors heavily concentrated in C2 simplification. Nasal assimilation and resyllabification were entirely absent in both varieties.

Pairwise Fisher's exact tests with Bonferroni correction (adjusted $\alpha=.005$ for 10 comparisons) revealed consistent patterns across both varieties. C2 simplification was the dominant subtype in both Seoul (19.2%, N=39) and Gyeongsang (15.6%, N=33), occurring significantly more frequently than C1 simplification (Seoul: $p<.001$; Gyeongsang: $p<.001$) and aspiration merger (both $ps<.001$). C1 simplification (Seoul: 8.9%, N=18; Gyeongsang: 4.7%, N=10) and aspiration merger (Seoul: 4.4%, N=9; Gyeongsang: 2.8%, N=6) did not differ significantly from each other in either variety (Seoul: $p=.083$; Gyeongsang: $p=.413$), though in Seoul the difference approached a marginal trend. Aspiration merger was significantly more frequent than nasal assimilation and resyllabification in Seoul (both $ps=.003$) but only marginally so in Gyeongsang (both $ps=.027$, uncorrected), falling just above the Bonferroni-corrected threshold.

When comparing Seoul and Gyeongsang directly for each individual subtype, none of the between-variety differences reached significance (C2 simplification: $p=.365$; C1 simplification: $p=.117$; aspiration merger: $p=.438$). The significant overall Seoul–Gyeongsang difference in phonologically-motivated errors therefore reflects a consistent directional pattern across subtypes rather than a single dominant subtype driving the effect.

6. Discussion

In this study, we examined the rates and patterns of errors produced by Whisper automatic speech recognition systems, focusing on Korean syllable-final consonant clusters and how these patterns vary across regional dialects, model sizes and Korean morphophonological processes. Below, we revisit the three predictions stated in Section 3 in light of the results of this study.

Prediction 1 (Dialect and model sizes): Both varieties show substantial improvement with model size, with error rates roughly halving from small to large-v3—a trajectory consistent with the general finding that larger Whisper models internalize Korean phonotactic patterns more robustly. While Gyeongsang consistently shows slightly higher error rates than Seoul at every model size, the gap narrows considerably across the scale and is effectively closed by the large-v3 model. Notably, the two varieties differ in where their gains are concentrated: Seoul's improvement is grounded in the reduction occurring between the small and medium models and little change thereafter, whereas Gyeongsang continues to improve more steadily across the larger model sizes. This asymmetry suggests that Seoul speech, which more closely matches the Seoul-standard training data, reaches a performance ceiling relatively early, while Gyeongsang speech—with its distinct phonological patterns—requires greater model capacity to be handled with comparable accuracy. The near-convergence of the two varieties at the large-v3 level is therefore not simply a product of general scaling but reflects the particular sensitivity of larger models to dialectal variation in consonant cluster contexts.

Prediction 2 (Influence of Korean morphophonological processes): The finding that phonologically-motivated errors constitute a substantial and consistent minority of all errors—ranging from 23.1% to 41.0% across varieties and models—confirms that Whisper's consonant cluster errors are not random but systematically reflect Korean phonological variations. In particular, C2 simplification was the dominant subtype in every model and variety, significantly exceeding all other subtypes. This dominance aligns with the recent findings that younger speakers favor the simplification of C2 over C1, and more broadly with the Seoul-standard training data in which C2-retaining forms predominate (Cho &Kim, 2009; Kim & Kang, 2021; Kwon et al., 2025). When Whisper fails to recover a full consonant cluster, it effectively defaults to the most typologically frequent surface outcome, suggesting that its language model has internalized Seoul-standard phonotactic tendencies.

The second tier of phonologically-motivated errors—C1 simplification and aspiration merger—behaved as a non-differentiated intermediate category across all models and varieties. The relatively high proportion of aspiration merger errors is somewhat surprising, given that aspiration merger applies obligatorily and exhibits no dialectal or speaker variation. By contrast, nasal assimilation and resyllabification—which are equally obligatory and uniform—were at floor throughout, indicating that Whisper handles these processes with considerably greater accuracy. The asymmetry between these obligatory processes suggests that the source of difficulty for aspiration merger lies not in phonological variability per se, but in the relative weakness of acoustic cues for /h/ in syllable-coda

position. Since /h/ contributes to surface form primarily through indirect effects such as aspiration of a following stop, the segmental information available to the model may be systematically underspecified. This points to a specific and tractable area for improvement: targeted fine-tuning with training data that richly represents aspiration merger contexts could substantially reduce this error type without requiring broader dialectal augmentation.

One finding that cuts across both predictions concerns the dialectal asymmetry in C1 simplification. Seoul showed a significantly higher rate of C1 simplification errors than Gyeongsang in the small model, a trend that attenuated progressively across larger models. This pattern runs counter to the initial expectation that Gyeongsang would produce more errors overall. A likely explanation lies in the interaction between Whisper's language model and ongoing phonological change in Seoul Korean: Seoul speakers' increasingly prevalent C1-retaining productions—well documented in recent corpus work (Cho & Kim, 2009; Kwon et al., 2025)—deviate from the Seoul-standard orthographic norm that dominates Whisper's training data, making them particularly susceptible to being misrecognized as C2-retaining forms. Gyeongsang speakers' C1-retaining productions, by contrast, are phonologically more categorical and may be more consistently mapped to recognizable surface patterns. The cross-model attenuation of this asymmetry further suggests that higher-capacity models gradually accommodate this ongoing change, though the mechanism—whether through broader phonetic coverage or more flexible language modeling—remains to be determined.

Additionally, Whisper's decoding leverages contextual language-model information, so frequently occurring lexical items may bias the system toward lexically plausible spellings, sometimes yielding over-restoration or substitution when acoustic evidence is compatible with multiple analyses (Radford et al., 2023). This lexical-bias interpretation is inferential here, since the present analysis does not explicitly model token frequency.

Future work should therefore (a) isolate cases where the same stem or morpheme is preserved to better approximate "pure" cluster processing, and (b) enrich the dataset with pronunciation and morpheme-boundary annotations so that resyllabification-eligible contexts can be directly conditioned.

7. Conclusion

This study presents a systematic, large-scale analysis of Whisper ASR errors in Korean consonant cluster recognition, using 54,536 spontaneous-speech utterances from the NIKL Daily Conversation Corpus 2022 (Version 1.0; Language Information Sharing Service '모두의 말뭉치'; NIKL, 2023) across two regional categories and four Whisper model variants. Overall error rates ranged from 2.67% (large-v3) to 4.53% (small). By contrast, the consonant cluster error rate ranged from 2.94% (large-v3, Seoul) to 6.93% (small, Gyeongsang).

Three key findings emerge from this study. First, the Seoul–Gyeongsang comparison reveals domain-specific dialect effects: while Seoul shows consistently higher overall coda accuracy, there is also interesting dialect-specific effect within phonologically-motivated errors: Seoul vs. Gyeongsang asymmetry in C1 simplification, which is significant under the small model and attenuates with increasing model size—a pattern we attribute tentatively to an interaction between Whisper's Seoul dialect-dominant language model and ongoing C1-retention change in Seoul Korean. Second, the error rates decrease with model size, with the mean error rate showing a great reduction from the smallest to the largest model variant. Accompanying this improvement is a narrowing of the Seoul–Gyeongsang gap in error rates, suggesting that model capacity is a meaningful moderator of dialect robustness. Importantly, this scaling benefit was not uniform across error types: phonologically motivated errors declined disproportionately faster than spurious errors, indicating that larger models have internalized Korean morphophonological alternations to a greater degree, while spurious errors proved comparatively resistant to scaling. Third, phonologically-motivated errors constitute a consistent and substantial minority (23%–41%) of all consonant cluster errors across models and varieties, with a statistically robust hierarchy placing C2 simplification as the dominant subtype. This suggests that Whisper's errors are not random but systematically reflect Korean morphophonological grammar, with the error hierarchy closely mirroring the typological frequency of phonological processes in Korean. The fact that such structure persists across all model sizes and both dialects indicates that these are not incidental confusions, but principled failure points rooted in the interaction between Korean phonotactics and Whisper's Seoul-dominant training data. This, in turn, suggests that incorporating explicit linguistic knowledge—such as phonotactic constraints, morphophonological rules, and dialect-specific alternation patterns—into training or post-processing pipelines may prove a more targeted and efficient strategy for reducing consonant cluster errors than scaling alone.

These findings carry both theoretical and practical implications. Theoretically, they confirm that ASR errors in morphophonologically complex languages are not arbitrary but reflect the structured grammar of the target language: the error hierarchy—with C2 simplification dominant, followed by C1 simplification and aspiration merger, and obligatory processes such as nasal assimilation and resyllabification at floor—closely mirrors the variable patterns of Korean phonological processes. Practically, the study moves beyond aggregate word error rate by identifying specific subtype-level vulnerabilities that can guide targeted intervention. C2 simplification errors—the dominant failure type—suggest that training data augmentation specifically representing consonant cluster environments and their surface alternations would yield more efficient gains than general-purpose scaling. Aspiration merger errors, despite their obligatory and invariant phonological status, emerged as a persistent source of difficulty, pointing to the acoustic underspecification of /h/ in coda position as a tractable target for fine-tuning. The asymmetric scaling trajectories of the two varieties further suggest that dialect-aware training—particularly for Gyeongsang speech—remains necessary even at the largest model sizes currently available.

Limitations of the present study include the absence of morpheme-boundary and following-segment annotations, which prevents systematic conditioning of errors by morphological environment; the restriction of evaluation to a single ASR system (Whisper), without direct comparison to other Korean ASR architectures; the use of orthographic rather than acoustic

evaluation, which prevents direct attribution of error patterns to phonetic cues such as coda release, aspiration, or gestural overlap; and the lack of control for sociolinguistic speaker variables (age, gender, education) that are known to interact with ongoing change in liquid-initial clusters. Addressing these limitations will enable a more causal account of how phonological variation, dialectal diversity, and model architecture jointly determine ASR behavior in morphophonologically complex languages like Korean.

In conclusion, this study demonstrates how a specific phonological domain—consonant clusters—can systematically shape ASR transcription under orthographic evaluation, and how this vulnerability is modulated by model size and dialectal variation. Future work should integrate morpheme-boundary, following-segment, and prosodic-boundary annotations, compare multiple ASR systems, and combine dialect- and variant-aware fine-tuning with acoustic and articulatory analyses to better identify the sources of ambiguity that drive consonant cluster errors in Korean.

Acknowledgement

This research was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea. Grant NRF-2021S1A5A2A01061716 was awarded to the first author, and Grant NRF-2022S1A5A2A01038289 was awarded to the corresponding author.

References

Cho, T., & Kim, S. (2009). Statistical patterns in consonant cluster simplification in Seoul Korean: Within-dialect interspeaker and intraspeaker variation. *Phonetics and Speech Sciences*, *1*(1), 33-40.

Cho, Y. Y. (1988). Phonological evidence for the lexical treatment of Korean suffixes. *Paper presented at the Annual Meeting of the Linguistic Society of America*. New Orleans, LA.

Cho, Y. Y. (2016). Korean phonetics and phonology. In *Oxford Research Encyclopedia of Linguistics*. Oxford, UK: Oxford University Press.

Jun, J. (1998). Restrictions on consonant clusters. *Linguistics*, *23*, 189-204.

Kim, J. Y. (2022). Variation in stem-final consonant clusters in Korean nominal inflection. *Glossa: a Journal of General Linguistics*, *7*(1), 5784.

Kim, J., & Kang, E. (2021). Phonetic variation of Korean stem-final consonant clusters beginning with a liquid. *Studies in Phonetics, Phonology and Morphology*, *27*(2), 161-192.

Kim-Renaud, Y. K. (1974). *Korean consonantal phonology*. Seoul, Korea: Pagoda Press.

Kwon, S., Yoon, T. J., Oh, S., & Han, J. I. (2023, July). Lexical diffusion in progress in Korean consonant clusters: A corpus study. *9th International Conference on Phonology and Morphology (ICPM9)*, Seoul, Korea.

Kwon, S., Yoon, T. J., Oh, S., & Han, J. I. (2025, January). Analogical generalization in progress in stem-final consonant cluster simplification: Evidence from Seoul and Gyeongsang Korean. *2025 Annual Meeting of the Linguistic Society of America*, Philadelphia, PA.

National Institute of Korean Language. (2017). Standard Pronunciation Rules (Pyojun Bal-eumbeop). Notification of the Ministry of Culture, Sports and Tourism. Retrieved from https://korean.go.kr/kornorms/regltn/regltnView.do?regltn_code=0002

National Institute of Korean Language. (2023). NIKL Daily Con versation Corpus 2022 (Version 1.0) [Data set]. Language Information Sharing Service (Modu Corpus). Retrieved from https://corpus.korean.go.kr/

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, Honolulu, HI.

Sohn, H. M. (1999). *The Korean language*. Cambridge. UK: Cambridge University Press.

Whitman, J. (1985). Korean clusters. In Kuno, S., Whitman, J., Lee, I. H., & Kang, Y. S. (Eds.), *Harvard studies in Korean linguistics* (pp. 280-290). Cambridge, MA: Harvard University.

• **Tae-Jin Yoon,** First author
Professor, Dept. of English Language and Literature, Sungshin Women's University
2, 34 da-gil, Bomun-ro, Sungbuk-gu, Seoul 02844, Korea
Tel: +82-2-920-7185
Email: tyoon@sungshin.ac.kr
Areas of interest: Acoustic phonetics, Speech Technology

• **Soohyun Kwon**, First author
Assistant Professor, School of Global Communication, Kyung Hee University
1732, Deogyeong-daero, Giheung-gu, Yongin-si, Gyeonggi-do 17104, Korea
Tel: +82-31-201-2291
Email: soohyunkwon@khu.ac.kr
Areas of interest: Sociophonetics, Phonological variation and change, Speech Techonology

• **Jeong-Im Han,** Corresponding author
Professor, Dept. of English Language and Literature, Konkuk University
120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Korea
Tel: +82-2-450-3339
Email: jhan@konkuk.ac.kr
Areas of interest: Acoustic phonetics, L2 acquisition