pISSN 2005-8063
eISSN 2586-5854
2025. 12. 31.
Vol.17 No.4
pp. 45-53

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2025.17.4.045

The effect of mismatched acoustic-articulatory information on listener judgments: The case of Korean cluster /lp/*

Sujin Oh[1,**] · Harim Kwon[2]

[1]*College of Humanities, Seoul National University, Seoul, Korea*

[2]*Department of English Language and Literature, Seoul National University, Seoul, Korea*

Abstract

This study investigates how listeners perceive the presence or absence of the stop consonant /p/ in the Korean /lp/ cluster, focusing on the relationship between acoustic and articulatory cues. Using data from 19 native speakers in the Seoul National University Multilingual Articulatory Corpus, we measured acoustic stop closure duration and lip aperture, and collected perceptual judgments from five native Korean listeners. Results revealed that acoustic and articulatory cues for /p/ strongly aligned for most speakers, with most tokens judged as having /p/ when both cues were present. Mixed-effects modeling showed that acoustic closure was the primary predictor of categorical /p/-identification, while articulatory lip closure alone had little effect. Crucially, several tokens exhibited mismatches between acoustics and articulation, showing either lip closure without acoustic silence interval or acoustic silence without lip contact. Listener judgments for these mismatched tokens were inconsistent and received lower goodness ratings, indicating perceptual uncertainty. These findings suggest that acoustic evidence alone does not always reflect articulatory events, highlighting the importance of incorporating both acoustic and articulatory information in speech perception research, especially when acoustic cues alone may not capture gestural nuances.

Keywords: acoustic-articulatory mismatch, speech perception, multimodal integration, listener judgment, Korean /lp/ cluster, cluster simplication

1. Introduction

Speech perception is inherently multimodal. Listeners integrate auditory information with other sensory cues relevant to articulation. Classic demonstrations such as the McGurk effect (McGurk & MacDonald, 1976) reveal that incongruent auditory-visual signals can alter perceptual outcomes, highlighting the non-trivial role of articulatory information in speech perception. More recent works

* This study is supported by the Creative-Pioneering Researchers Program at Seoul National University awarded to Harim Kwon.

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

Received 4 November 2025; Revised 9 December 2025; Accepted 10 December 2025

(e.g., Fowler & Dekle, 1991; Gick & Derrick, 2009) show that even subtle tactile cues about articulation can also influence listeners' judgments. For example, Gick & Derrick (2009) demonstrated that slight, aspiration-like air puffs applied to the skin increase the likelihood that listeners categorize stops as aspirated, indicating that somatosensory information is recruited as event-relevant perceptual evidence during speech perception. Such findings suggest that speech perception is guided by articulatory gestures assessed through multiple sensory channels. Listeners actively integrate multiple sources of evidence, including, but not limited to acoustic evidence, to extract stable linguistic information from often ambiguous inputs.

Studies have shown that listeners do not rely solely on the acoustic signal but draw on expectations about articulation. For example, studies on compensation for coarticulation demonstrate that listeners adjust their perception based on expectations about overlapping gestures (e.g., Fowler, 1981; Mann & Repp, 1980). Listeners can factor out coarticulatory effects when certain articulatory configurations are plausible within the phonological context (e.g., Manuel, 1995; Warner & Weber, 2001). These findings suggest that mismatches between acoustics and underlying articulatory events may not necessarily map transparently onto perception, motivating the central question of the present study.

Experimental work in phonetics and phonology often infers that the absence of an acoustic cue (e.g., a silence interval for stop closure) reflects the absence of the corresponding articulatory gesture, and by extension, speaker intent. This inference follows from the general view that acoustic signals arise as consequences of articulatory movements, but it is not a necessary correlation. Nevertheless, it is not universally the case. For example, articulatory gestures can occur without clear acoustic correlates, as in the classical example of 'perfect memory.' When produced casually, the final /t/ in the word 'perfect' is articulatorily produced but its acoustic consequences can be completely hidden as the tongue tip gesture overlaps with adjacent velar and labial gestures (e.g., Browman & Goldstein, 1991). Conversely, apparent acoustic events may exist without articulatory evidence, as in the case of acoustic schwas without a corresponding vowel gesture due to decreased overlap (e.g., Browman & Goldstein, 1990, 'beret' vs. 'bray') or articulatory mistiming (e.g., Davidson, 2005). Such discrepancies raise important questions: How do mismatches between acoustic and articulatory evidence influence listeners' judgments?

To address this question, we examine a naturally variable case, consonant cluster simplification in Korean. Korean maximal syllable structure is C(G)VC, and thus consonant clusters in coda positions are typically simplified through re-syllabification or deletion, depending on their phonological contexts. For instance, the /ps/ cluster in <갑> /kaps/ 'price' is simplified through re-syllabification when the vowel-initial suffix /-ɨl/ is added as shown in example (1). In isolation or when followed by a consonant-initial suffix, the cluster is simplified through deletion of /s/, as in examples (2) and (3).

/kaps + ɨl/ 'price + ace.' → [kap.sɨl] (1)
/kaps/ 'price' → [kap] (2)
/kaps + to/ 'price + as well' → [kap.t*o] (3)

Although this tendency of preserving C_1 and deleting C_2 is observed across various cluster types, the /lp/ clusters, which serve as the test case for the current investigation, are reported to be the most variable realizations among other types of consonant clusters in Korean (e.g., Kim, 2022; Kwon et al., 2023, 2025; Nam & Oh, 2009; Yun, 2023). For example, <넓다> /nəlp.ta/ 'wide' may surface as either [nəl.t*a] or [nəp.t*a]:

/nəlp + ta/ 'wide + dec.' → [nəl.t*a]~[nəp.t*a] (4)

Previous studies on this variation report a preference for /l/-realization in the /l/+stop clusters, including both /lp/ and /lk/ clusters. For example, Nam & Oh (2009) and Kim (2022) report that speakers of the Seoul dialect tend to simplify /lp/ to [l], irrespective of the following segment. Similarly, Kwon et al. (2023, 2025) found a stronger tendency for deletion of /p/ in the /lp/ cluster than for /k/ in the /lk/ cluster, implying the preference to /p/-deletion in the /lp/ cluster. Though these studies commonly suggest a preference for /l/ retention (and /p/ deletion) for /lp/ clusters, they also have a common limitation. That is, they relied exclusively on acoustic analysis and listener judgments to categorize the actual production of the /lp/ clusters. However, as mentioned above, the alignment between acoustics and articulation (and speaker intent) is not guaranteed in all contexts. Therefore, the absence of acoustic or perceptual evidence should not be taken as direct evidence for phonological deletion or the absence of articulatory effort.

While articulation has long been disregarded in this line of research, a recent study by Yun (2023) uses ultrasound tongue imaging to investigate whether the /l/+stop cluster simplification is categorical (i.e., speakers intended to delete the segment) or gradient (i.e., speakers made the articulatory efforts, but the gestures were reduced). Based on the tongue images of both /lp/ and /lk/ clusters, they claim that the /lk/ cluster exhibited a gradient reduction of /k/ while the /lp/ cluster was categorized into two groups: deleting /l/ or /p/. While Yun (2023) made a valuable contribution by moving beyond acoustics and integrating articulatory data, this method remains limited in the case of /p/, which is a non-lingual consonant. The labial constriction for /p/ cannot be adequately captured by tongue imaging as it does not necessarily require a specific tongue position. Determining whether /p/ was retained or deleted can only be properly addressed by examining lip articulation. In addition, it still remains unanswered how the lip articulation (especially when it does not align with the acoustics) contributes to the listeners' perceptual judgments.

To address this gap, we examine two central questions:

1. Does the absence of acoustic closure for /p/ necessarily imply the absence of a corresponding lip closure gesture?
2. When acoustic and articulatory cues diverge, how does the mismatch influence listeners' categorical judgments and their goodness ratings?

We aim to clarify how acoustic and articulatory information interact in guiding perception and what mismatches reveal about the robustness of cue integration in speech processing. Crucially, we do not intend to provide a phonological description of Korean clusters in general, such as the phonological classification of /p/ deletion or its presence within the cluster, as investigated in the previous study above. Instead, we leverage the variable realizations of the /lp/ cluster as an empirical context in which the acoustic and articulatory evidence for /p/ may diverge. By targeting a single lexical item

containing /lp/, we test how fine-grained articulatory (lip closure) and acoustic (silence interval) information influence listener perception.

2. Methodology

2.1. Speech Materials

We investigated the lip closures associated with the /p/ phoneme in /lp/ clusters, using the data from Seoul National University Multilingual Articulatory Corpus (SNU-MAC; Kwon & Oh, 2025). This corpus comprises ultrasound tongue imaging, lip videos from both front and side perspectives, and corresponding audio recordings, simultaneously recorded. We selected lip front videos and audio recordings from 19 native Korean speakers, comprising 16 females and three males (Mean age=29.37; SD=8.45; range=19~52). Most speakers (14 out of 19) were from the Seoul/Gyeonggi dialect region, and the rest were from Gangwon, Gyeongsang, or Jeolla. All participants reported normal hearing and no speech or language disorders.

The target word is '얇습니다' /jalpsɨpnita/, which contains the /lp/ cluster followed by the fricative /s/. The word was selected because it is the sole word within the controlled stimuli corpus that contains the /lp/ cluster.

2.2. Measurements

Since we primarily focus on identifying the acoustic and articulatory evidence for the /p/ in the /lp/ cluster, we conducted two measurements: An acoustic measure for the silence interval duration (StopDuration) and an articulatory measure to examine the lip closure (LipDistance). The synchronization between the two signals (lip front video recording and acoustic recording) was achieved in Articulated Assistant Advanced software (Articulate Instrument, 2012) based on the sync signals in both recordings. The time-synched signals were subsequently subjected to acoustic and articulatory analyses, respectively. The elicited target word '얇습니다' (four syllables) had an average length of 818 ms, with a standard deviation of 113 ms.

For StopDuration, we measured the duration of the silence interval for /p/ between /l/ and /s/ (in milliseconds) based on waveforms and spectrograms in Praat (Boersma & Weenink, 2025). Specifically, we measured the silence interval from the end of the regular periodicity of the preceding sonorant (either /l/ or the preceding vowel /a/, if the speaker deleted /l/ in /jalpsɨpnita/) to the onset of the fricative noise in the following /s/. When the offset of the sonorant was unclear, we relied on the presence of higher formants (above F2) to determine the offset. If no higher formants were visible, we treated the interval as silence.

For LipDistance, we concentrated on the lip constriction and measured the distance between the upper and lower lips in the frames where a bilabial constriction was expected (Yip, 2013). First, we extracted the frames between the preceding /l/ or /a/ and the following /s/ from the lip video recordings. Then, as depicted in Figure 1, the spatial points were directly drawn onto the corresponding frame images for the upper and lower lips, represented by blue and red dots, respectively. These two points were put at the lower edge of the upper lip and the upper edge of the lower lip. The distance between these two points was measured to determine the bilabial constriction. The minimal distance value (in millimeters) for each token was labeled as LipDistance.

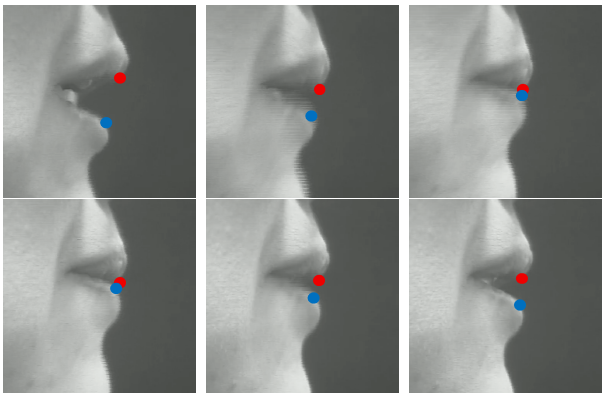

Figure 1. Example frames from lip video showing measurement of lip aperture. Blue and red dots mark upper and lower lip edges, respectively. The distance between these points quantifies lip closure during /p/ production.

2.3. Listener judgments

To assess listeners' perception of the presence/absence of the /p/ in the /lp/ cluster, we conducted a listener judgment task. Five linguistically trained native Korean listeners were presented with audio recordings of the target word and tasked with determining whether the target word is produced with one of the two consonants or with the full cluster (i.e., C_1, C_2, or C_1C_2). Additionally, they were asked to rate each token using a 5-point Likert-like scale (1='poor', 5='excellent') to assess how good the selected label matches the sound they heard.

2.4. Analysis

All the analysis was conducted in R version 4.5.1 (R Core Team, 2025). The packages used in the analysis are car (Fox & Weisberg, 2018), effects (Fox & Weisberg, 2018), emmeans (Lenth, 2025), irr (Gamer et al., 2019), lme4 (Bates et al., 2015), ordinal (Christensen, 2023), and performance (Lüdecke et al., 2021).

2.4.1. Coding

A total of 19 tokens, each produced by different speakers, were rated by five listeners. The resulting dataset comprised 19 acoustic StopDuration and articulatory LipDistance measurements and 95 listener judgments. To classify each token as having versus not having acoustic, articulatory, and perceptual evidence, the measures and listener judgment responses were converted to binary indicators for /p/ (present/absent) in the following manner.

Gaussian mixture modeling (GMM) was applied separately to StopDuration and LipDistance measures to identify natural clustering corresponding to "presence" versus "absence" of /p/. Prior inspection of the distributions showed a clear bimodal tendency for both variables, suggesting two underlying categories (see Figure 2). For each acoustic and articulatory measure, a two-component GMM provided the best fit based on the Bayesian information criterion (BIC). For StopDuration, the component with higher mean was labeled as acoustic stop present; for LipDistance, the component with lower mean (smaller aperture) was labeled as lip closure present. These yielded two binary variables: AcousticStop (1=acoustic evidence of /p/ present; 0=absent) and

LipClosure (1=lip closure present; 0=absent). Continuous predictors were mean-centered and scaled for modeling.

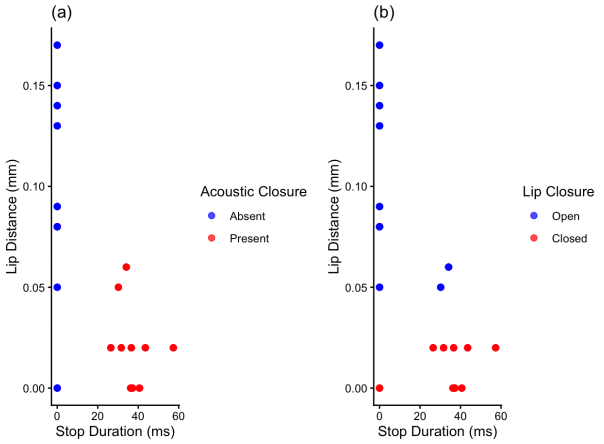

Figure 2. Two-dimensional scatter plots of StopDuration and LipDistance classified using Gaussian mixture models. Panel (a) shows acoustic closure classification, and Panel (b) shows lip closure classification.

Listeners' categorical judgments of /p/ presence were binarized into ListenerJudgment (1=perceived /p/ present, 0=absent). That is, both the C_2 responses and the C_1C_2 responses were coded as 1, while the C_1 responses were as 0. Goodness ratings were treated as ordinal (1–5).

2.4.2. Statistical Analysis

The analysis was conducted based on the two central questions outlined in Section 1.

RQ1: Agreement between acoustic and articulatory evidence was assessed via Cohen's kappa and McNemar's test on the binary indicators (AcousticStop, LipClosure). Additionally, the continuous measures, StopDuration and LipDistance, were examined using linear regression to assess whether the two measures correlate with each other. Logistic regression models predicting binary LipClosure from continuous StopDuration and binary AcousticStop from LipDistance were also fitted to estimate the relation between acoustics and articulations.

RQ2: A generalized linear mixed-effects model (GLMM) with a logit link was used to predict ListenerJudgment from both binary indicators and their interaction: AcousticStop, LipClosure, and AcousticStop×LipClosure, with random intercepts for SpeakerID and ListenerID to account for repeated listener judgments and ratings per token. Goodness ratings were modeled with a cumulative link mixed model (CLMM) using the same predictors and random effect structure to respect the ordinal nature of the scale (Taylor et al., 2023). All models included centered continuous covariates when applicable, and model assumptions (e.g., linearity in logits, proportional odds for CLMM) were evaluated.

3. Results

Table 1 provides descriptive statistics for all raw measures, including StopDuration, LipDistance, and the listener responses (judgments and ratings).

The results from the listener judgment task revealed that tokens categorized as C_1 were 38.95%, tokens judged as C_1C_2 were 55.79%, and tokens as C_2 were only 5.26%. In total, 61% of the tokens were judged as having /p/.

Table 1. Descriptive statistics for acoustic stop duration (ms), lip distance (mm), and listener goodness ratings across all tokens

Perceived category	N	Percent (%)	Stop duration (ms)		Lip distance (mm)		Goodness rating	
			M	SD	M	SD	M	SD
C1	37	38.95	0.92	5.61	0.11	0.05	3.22	0.89
C1C2	53	55.79	31.50	16.40	0.03	0.03	2.92	0.98
C2	5	5.26	33.50	2.54	0.01	0.01	3.40	0.89

M, mean; SD, standard deviation.

3.1. Relation Between Acoustic and Articulatory Cues

Table 2 presents the cross-classification of 19 tokens according to the acoustic and articulatory evidence for /p/. The two indicators aligned for most cases, with only three tokens of mismatch. To statistically investigate the relation between the acoustic and the articulatory evidence for /p/, we first assessed the agreement between the two binary indicators (i.e., AcousticStop, LipClosure) using Cohen's kappa and McNemar's test with continuity correction. The Cohen's kappa indicated a substantial level of agreement between the two modalities (κ =0.69, p=.002). McNemar's test also revealed no significant difference between the two modalities (χ^2=0, p=1).

Table 2. Contingency table showing agreement between binary indicators: Acoustic stop and lip closure

		Articulatory /p/		Total
		Absent	Present	
Acoustic /p/	Absent	8	1	9
	Present	2	8	10
Total		10	9	19

A Pearson correlation analysis was conducted to examine the relation between the two continuous measures, StopDuration and LipDistance. The results (see Figure 3) indicated a strong correlation, $r(17)$=–0.70, p<.001, suggesting that the shorter the stop silence duration is, the farther the distance between the upper lip and the lower lip is.

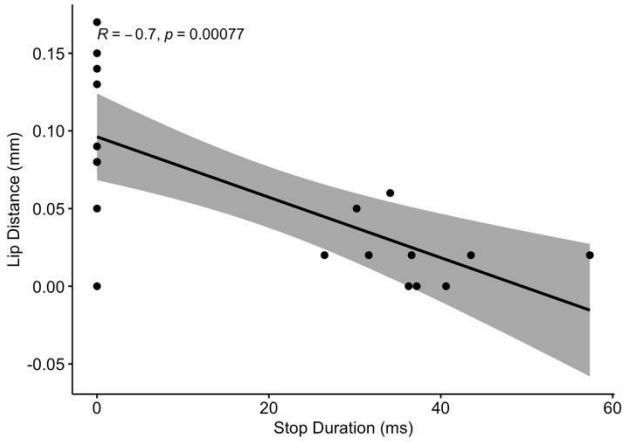

Figure 3. The correlation between lip distance and stop duration.

Finally, we fitted a logistic regression model to predict the binary LipClosure variable from the continuous StopDuration. The model showed that the likelihood of observing a lip closure increased with greater acoustic silence duration (β=.102, *SE*=.039, *z*=2.61, *p*<.01). When fitting the same model predicting the binary AcousticStop variable from the continuous LipDistance, it shows that the likelihood of observing an acoustic closure increased with greater lip distance (β=−51.68, *SE*=23.01, *z*=−2.25, *p*<.05).

All three analyses we conducted to address RQ1 indicated a close relation between acoustics and articulation. However, despite the general agreement between acoustic and articulatory cues for /p/, some tokens did exhibit a discrepancy between these two modalities during the production of the /p/. As shown in Table 2, three speakers (out of 19, over 15%) manifested the acoustic-articulatory mismatch in their productions. The details of these mismatch cases and how they affect the listeners' judgments are examined in Section 3.3.

3.2. Predicting Listener Judgments

To address how listeners integrate acoustic and articulatory cues in their perceptual judgments—particularly under conditions in which the two cues mismatch—we fitted a generalized linear mixed-effects model (GLMM) to the binary ListenerJudgment responses. The model included fixed effects of the binary variables, AcousticStop and LipClosure, along with their interaction. It also included random intercepts for SpeakerID and ListenerID. All variance inflation factors (VIFs) for fixed effects were below 2.5, indicating no significant multicollinearity among predictors.

The GLMM revealed a significant main effect of acoustic closure on perceived stop identification (β=5.97, *SE*=2.46, *z*=2.43, *p*<.05), indicating that the presence or absence of acoustic closure substantially influenced whether listeners categorized the token as containing a stop. In contrast, the main effect of articulatory (lip) closure was not significant (β=0.97, *SE*=2.83, *z*=0.34, *p*=.73), suggesting that visual articulatory information alone had little impact on categorical stop judgments (see Figure 4). The interaction between acoustic and articulatory cues was also non-significant (β=18.30, *SE*=915.89, *z*=.02, *p*=.999), providing no evidence that cue mismatch altered the effect of either cue on categorical judgments. Note, however, the interaction term yielded an unstable estimate with an extremely large standard error, likely due to sparse mismatch cases. Given the uncertainty, this null result should be interpreted cautiously.

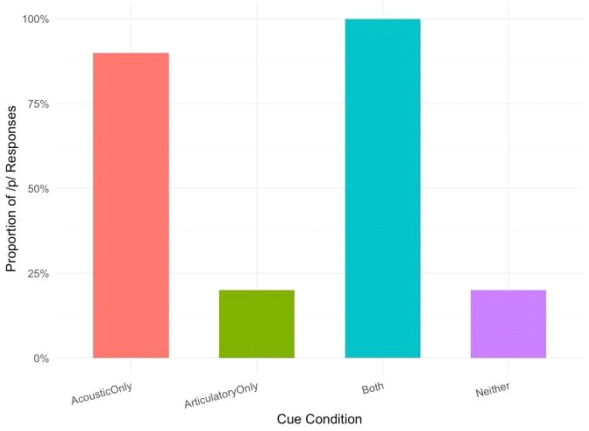

Figure 4. Proportion of /p/ judgments across cue conditions.

For the goodness rating responses, we fitted a CLMM with the same fixed and random structure. The acoustic cue showed a marginal effect (β=−1.61, *SE*=0.92, *z*=−1.74, *p*=.082), with tokens lacking acoustic closure tending to receive lower goodness ratings, though this effect did not reach conventional significance (Figure 5). The articulatory cue again showed no significant effect (β=−0.86, *SE*=1.26, z=−0.68, *p*=.49), and the interaction between the two cues was non-significant (β=2.61, *SE*=1.58, *z*=1.65, *p*=.099).

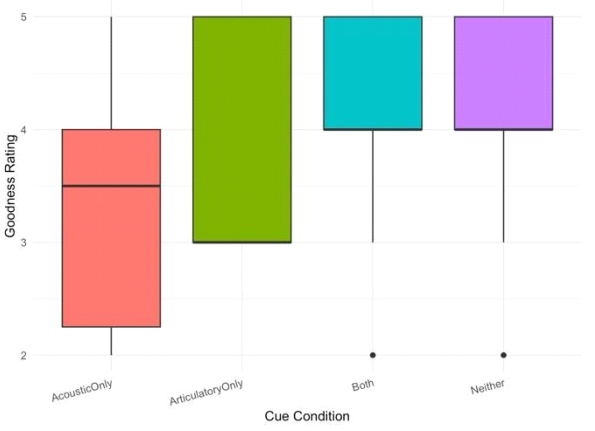

Figure 5. Goodness ratings by cue presence.

3.3. Mismatch Analysis

To better understand the roles of the discrepancy between acoustic and articulatory cues in the perception of the /p/ in the /lp/ cluster, the mismatch cases were individually examined.

Table 3 showed that 19 tokens produced by 19 speakers are divided into four conditions based on the acoustic and articulatory measurements.

Table 3. The acoustic and articulatory measurements by condition

Condition	N	Stop duration (ms)		Lip distance (mm)	
		M	SD	M	SD
Acoustic-only	2	32.20	2.06	0.06	0.01
Articulatory-only	1	0.00	0.00	0.00	0.00
Both	8	38.70	8.67	0.01	0.01
Neither	8	0.00	0.00	0.11	0.04

N, the number of spoken tokens; M, mean; SD, standard deviation.

Most listeners reported not perceiving /p/ when both articulatory and acoustic evidence for the segment was lacking (i.e., Neither condition in Table 3). Out of the 40 perceptual judgments (8 tokens×5 listeners) in this condition, 32 (80%) were judged as not having /p/ (i.e., /l/-only), indicating that the congruent information elicited highly concordant responses for most listeners and most tokens (see Figure 4). Moreover, the goodness ratings for these tokens were relatively high, averaging 4.28 out of 5.0, indicating that listeners found the correspondence between the label and the sound to be clear and reliable (see Figure 5). Figure 6 visually illustrates the alignment of acoustic signals (waveform and spectrogram in the top panel) and articulatory signals (lip distance in the bottom panel). This alignment demonstrates the absence of stop closure (in the highlighted part) during the production of the segment of the target word, /jalps.../.

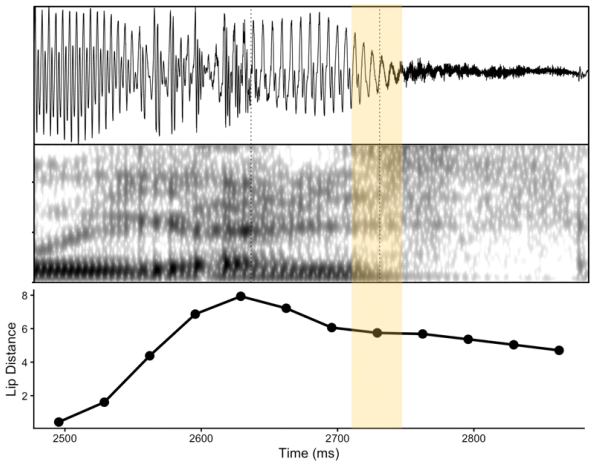

Figure 6. The acoustic-articulatory match in the production of /p/: The case of Neither (talker010).

When both acoustic and articulatory evidence are present (i.e., Both condition), listeners unanimously perceived the /p/ in the clusters. All 40 perceptual judgments (100%) identified the presence of /p/ (i.e., giving either /p/ or /lp/ responses) when the two cues were matched (see Figures 4 and 7). These tokens also received high ratings (*M*=4.2), suggesting a strong perceived fit between the label and the sound (also see Figure 5). Figure 7 visually demonstrates the alignment of the acoustic and articulatory signals during the stop closure.

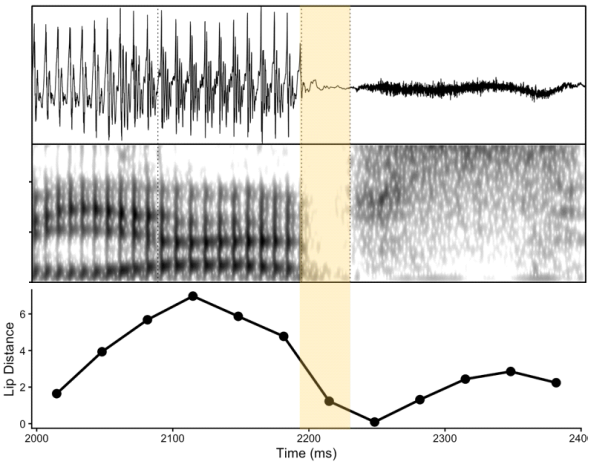

Figure 7. The acoustic-articulatory match in the production of /p/: The case of Both (talker014).

Unlike the two matching conditions, listeners' responses diverged in the three cases with a mismatch between acoustic evidence and articulatory evidence for /p/. Out of three, two speakers failed to exhibit clear lip closure for /p/, yet they produced acoustic silence intervals (mean duration=32.2 ms). However, the lip distance never reached 0 mm, indicating that the upper and lower lips did not fully close during the /lp/ production.

Figure 8 visually demonstrates the mismatch between two pieces of evidence from one of the two speakers' productions. The temporal region where the waveform and the spectrogram indicate a stop silence interval is highlighted. The same region in the bottom panel indicates the lips did not reach its full closure during this interval. The later lip closure (around 2700 ms) reflects the /p/ in the following syllable /sɨp/, coproduced with /s/ and devoiced /ɨ/.

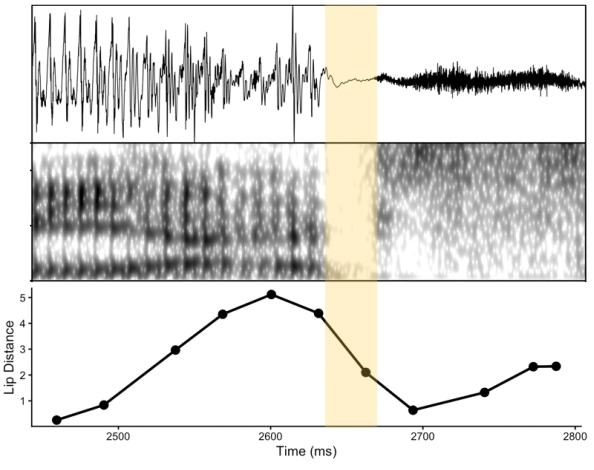

Figure 8. The acoustic-articulatory mismatch in the production of /p/: The case of Acoustic-only (talker023).

Out of ten perceptual judgments (2 tokens×5 listeners), nine judgments (90%) categorized the token as having /p/, indicating that the listeners perceived /p/ without the articulatory evidence. These acoustic-only tokens were also rated lower than the matched tokens, with an average rating of 3.44 (*SD*=1.24), suggesting that listeners found these tokens harder to categorize confidently.

On the contrary, one speaker demonstrated the opposite pattern—articulatory lip closure without a corresponding acoustic stop interval. Figure 9 illustrates the discrepancy. The temporal region where /p/ is expected (highlighted in yellow) does not exhibit a silence interval according to the formant-based definition of silence. However, the change in lip distance indicates that the lip distance reached its minimum (i.e., the threshold for the classification of lip constriction in GMM) for a short while, suggesting articulatory evidence for /p/ production.

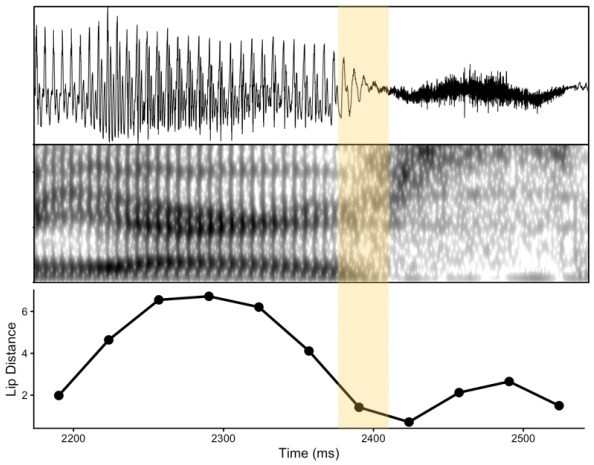

Figure 9. The acoustic-articulatory mismatch in the production of /p/: The case of Articulatory-only (talker020).

Out of five listener judgments, one (20%) classified the token as containing /p/, suggesting that the articulatory cue alone sometimes triggered a perception of /p/ in the absence of acoustic silence. This token was also rated relatively low, with the mean rating of 3.0, indicating a higher degree of uncertainty in the listener's judgments compared to the matched tokens. This suggests that mismatches between articulatory and acoustic cues can lead to perceptual uncertainty.

4. Discussion

The current study examined the relation between acoustic and articulatory cues and how the acoustic-articulatory mismatch influences listeners' judgments, using Korean /lp/ clusters as a test case. Our findings revealed that a strong overall alignment between acoustic and articulatory cues. For the majority of speakers, articular evidence of lip closure and acoustic evidence for stop silence interval co-occurred. This confirms the common assumption that acoustic and articulatory signals typically align, both reflecting speaker intent.

However, a closer look at the data shows that mismatches are not negligible. Three out of 19 speakers (over 15%) showed a mismatch between articulatory and acoustic information. In two cases, lip closure occurred without a corresponding acoustic silence interval, and in one case, an acoustic silence occurred without lip contact. It should be noted that such mismatches were far less frequent than matches and our sample size is limited, and that GMM-based binary classification may introduce some over- or under-classification, particularly in small samples where mixture-model cluster boundaries can be less stable. Still, the occurrence of mismatches in over 15% of speakers suggests that such cases may not be negligible and thus must be taken seriously in speech production and perception research.

Listener judgments further underscore the importance of these mismatches. When the two cues are congruent, perception was stable: listeners consistently reported the presence or absence of /p/ and assigned relatively high goodness ratings. However, mismatched tokens triggered perceptual uncertainty. Overall, the acoustic cues dominate categorical judgments (acoustic-only tokens were more frequently categorized as having /p/ than articulatory-only tokens), yet the mismatches introduce perceptual uncertainty. Although only a small number of tokens fall into this category, their impact on listener uncertainty suggests their theoretical significance/relevance.

The current findings challenge the common assumption that "no acoustic silence=no stop articulation." Our results show that articulatory evidence may still be present without acoustic silence, suggesting the acoustic-only analyses risk mischaracterizing speaker intent. This has both theoretical and methodological implications for studies in experimental phonetics/phonology, especially when acoustic analyses and perceptual coding are the primary sources of evidence. For example, in the case of Korean /lp/ clusters, previous studies have largely relied on linguistically-trained listeners' coding of acoustic recordings. The perceptual data in the present study closely resemble such methods: listeners frequently reported not hearing /p/ even when a lip closure gesture was present. This indicates that perceptual judgments do not always reflect the articulatory events produced by speakers, especially in mismatched tokens. Consequently, what has been labeled as /p/-deletion in earlier studies may not always be the case of an actual absence of /p/ articulation. This highlights the importance of incorporating articulatory evidence, pointing to the need for future studies to reassess deletion judgments that rely solely on acoustic or perceptual data.

In the context of sound change (as in the case of Korean coda clusters, see e.g., Kwon et al., 2025, for more discussion on this), the theoretical implications are even more significant. One of the theoretical accounts of sound change posits that it arises, at least in part, from the misalignment between the speaker's intent and the listener's interpretation. Such misalignments can stem from contextual variability, such as coarticulation, which obscures the intended target in systematic and lawful ways and leads to perceptual reanalysis (e.g., Beddor, 2009; Ohala, 1981). In our view, visually transmitted articulatory information is relevant to these misalignments, as such information may help listeners recover the speaker's intent (hence prevent perceptual reanalysis) particularly when the sound under change involves visually distinctive articulations (e.g., labial consonants or round vowels). Prior findings support this possibility showing that visual information is indeed used in identifying speech sounds involving labial articulations (Jongman et al., 2003; Stephens & Holt, 2010).

The importance of visual information is further supported by evidence from language acquisition. Infants attend more to speakers' mouth than to their eyes (Hillairet de Boisferon et al., 2018) and are highly sensitive to the lip gestures of caregivers, aligning the auditory and visual aspects of speech (e.g., Burnham & Dodd, 2004; MacKain et al., 1983; Rosenblum et al., 1997). If sound change is a listener-based, generational process, these findings highlight the need for researchers to incorporate visual

articulatory information into experimental work on sound change. Yet, most experimental approaches to sound change have overlooked the potential role of visible articulation in shaping perceptual interpretation and, ultimately, sound change. Taken together, these considerations suggest that ecologically valid experimental approaches to sound change should consider incorporating visible articulatory information, especially when the sounds under investigation are visually salient. Relying solely on the perceptual judgments based on acoustic signals risks researchers' misclassifying what speakers actually produce and what listeners actually perceive.

Furthermore, the mismatch cases highlight promising directions for future research. As the current study relied on audio-only perception, listeners had no access to visual information that might have increased the weight of articulatory cues (as in real-life conversation). That is, it is plausible that the mismatch tokens would be perceived differently if listeners had access to visual information. To address this, we are currently preparing follow-up experiments incorporating audio-visual stimuli to test whether seeing lip gestures indeed alter perception of articulatory-only tokens. Such work will clarify whether the perceptual asymmetry (more weight to acoustics) observed in the current study is a by-product of audio-only methods and whether access to visual information on articulatory gestures rebalance the cue weighting.

Finally, the current study has some limitations that point toward future research. The perceptual results are based on a small group of linguistically training listeners, making it difficult to generalize the findings. Trained listeners may attend more closely to certain cues than naïve listeners, potentially amplifying or even altering the perceptual patterns reported in this study. An important next step, therefore, is to examine whether similar perceptual patterns arise among naïve listeners with no phonetics or linguistic training, and whether the relative importance of acoustic versus articulatory cues changes as a result of linguistic training. Expanding listener sample in the follow-up study will help assess the generalizability of the findings and clarify how cue mismatches are perceived by naïve listeners.

5. Conclusion

The current study examined the relationship between acoustic and articulatory cues in listener judgments of /p/ presence/absence in Korean /lp/ cluster. Although the previous study suggested the deletion of /p/ in the in the cluster /lp/ by showing the preference to /l/ realization, the judgments were solely based on acoustic analysis and auditory perception. In contrast, our results demonstrate that articulatory evidence may be present even when no corresponding acoustic signal is detected, and this discrepancy can lead to uncertainty in listeners' judgments about whether a sound has been deleted. Thus, it is important to consider articulatory gestures when evaluating listener perceptions.

References

Articulate Instruments. (2012). *Articulate assistant advanced user guide: version 2.14*. Edinburgh, UK: Articulate Instruments Ltd.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Beddor, P. S. (2009). A coarticulatory path to sound change. *Language, 85*(4), 785-821.

Boersma, P., & Weenink, D. (2025). Praat: Doing phonetics by computer (version 6.4.38) [Computer program]. Retrieved from https://www.praat.org/

Browman, C. P., & Goldstein, L. (1991). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston, & M. E. Beckman (Eds.), Papers in laboratory phonology I: Between the grammar and the physics of speech (pp. 341-376). Cambridge, UK: Cambridge University Press.

Browman, C. P., & Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics, 18*(3), 299-320.

Burnham, D., & Dodd, B. (2004). Auditory–visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology, 45*(4), 204-220.

Christensen, R. H. B. (2023). Ordinal: Regression models for ordinal data, R package version 2023.12-4.1. Retrieved from https://CRAN.R-project.org/package=ordinal

Davidson, L. (2005). Addressing phonological questions with ultrasound. *Clinical Linguistics & Phonetics, 19*(6-7), 619-633.

Fowler, C. A. (1981). A relationship between coarticulation and compensatory shortening. *Phonetica*, *38*(1-3), 35-50.

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 17*(3), 816-828.

Fox, J., & Weisberg, S. (2018). *An R companion to applied regression* (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). irr: Various Coefficients of Interrater Reliability and Agreement, R package version 0.84.1. Retrieved from https://doi.org/10.32614/CRAN.package.irr

Gick, B., & Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature, 462*(7272), 502-504.

Hillairet de Boisferon, A., Tift, A. H., Minar, N. J., & Lewkowicz D. J. (2018). The redeployment of attention to the mouth of a talking face during the second year of life. *Journal of Experimental Child Psychology, 172,* 189-200.

Jongman, A., Wang, Y., & Kim, B. H. (2003). Contributions of semantic and facial information to perception of nonsibilant fricatives. *Journal of Speech, Language, and Hearing Research, 46*(6), 1367-1377.

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

Kwon, H., & Oh, S. (2025, May). The SNU multilingual articulatory corpus: A resource on cross-linguistic speech production research [conference presentation]. *The 2025 Joint Meeting of Korean Society of Speech Sciences, the Phonology-Morphology Circle of Korean, and the Korean Association for the Study of English Language and Linguistics*. Seoul National University, Seoul, Korea.

Kwon, S., Oh, S., Yoon, T., & Han, J. I. (2025, January). Analogical change in progress in the Korean consonant cluster simplification: a corpus study [conference presentation]. *The 2025 LSA Annual Meeting*, Philadelphia, PA.

Kwon, S., Yoon, T. J., Oh, S., & Han, J. I. (2023, May). Variable

realization of consonant clusters in Seoul and Gyeongsang Korean. *Proceedings of Hanyang International Symposium on Phonetics and Cognitive Sciences of Language* (pp. 58-59). Seoul National University, Seoul, Korea.

Lenth, R. (2025). emmeans: Estimated marginal means, aka least-squares means, R package version 1.11.2. Retrieved from https://doi.org/10.32614/CRAN.package.emmeans

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). Performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, *6*(60), 3139.

MacKain, K., Studdert-Kennedy, M., Spieker, S., & Stern, D. (1983). Infant intermodal speech perception is a left-hemisphere function. *Science, 219*(4590), 1347-1349.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Perception & Psychophysics, 28*, 213-228.

Manuel, S. Y. (1995). Speakers nasalize /ð/ after /n/, but listeners still hear /ð/. *Journal of Phonetics*, *23*(4), 453-476.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

Nam, K., & Oh, J. (2009). The analysis of the reasons and aspects of pronunciation of 'lk/lp'. *Korean Linguistics, 42*, 123-153.

Ohala, J. J. (1981). Articulatory constraints on the cognitive representation of speech. *Advances in Psychology, 7*, 111-122.

R Core Team. (2025). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics, 59*(3), 347-357.

Stephens, J. D. W., & Holt, L. L. (2010). Learning to use an artificial visual cue in speech identification. *The Journal of the Acoustical Society of America, 128*(4), 2138-2149.

Taylor, J. E., Rousselet, G. A., Scheepers, C., & Sereno, S. C. (2023). Rating norms should be calculated from cumulative link mixed effects models. *Behavior Research Methods, 55*(5), 2175-2196.

Warner, N., & Weber, A. (2001). Perception of epenthetic stops. *Journal of Phonetics*, *29*(1), 53-87.

Yip, J. C. K. (2013). *Phonetic effects on the timing of gestural coordination in Modern Greek consonant clusters* (Doctoral dissertation). University of Michigan, Ann Arbor, MI.

Yun, G. (2023). An articulatory study on consonant cluster simplification in L1 Korean and L2 English. *Korean Journal of English Language and Linguistics, 23*, 1169-1193.

● **Sujin Oh,** Corresponding author
Postdoctoral Researcher, College of Humanities, Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
Email: sujinoh@snu.ac.kr
Areas of interest: Phonetics, L2 acquisition

● **Harim Kwon**
Associate Professor, Dept. of English Language and Literature, Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
Tel: +82-2-880-6112
Email: harimkwon@snu.ac.kr
Areas of interest: Phonetics, Laboratory phonology