Korean Journal of English Language and Linguistics, **Vol 26, February 2026, pp. 275-285**
DOI: 10.15738/kjell.26..202602.275

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

Measuring Vowel Contrast Revisited*

Ye-Yeong Han · Joo-Kyeong Lee (University of Seoul)

Received: September 24, 2025
Revised: December 26, 2025
Accepted: February 2, 2026

Han, Ye-Yeong (First author)
MA student, Department of English Language and Literature
University of Seoul
Email: ye0626ye@gmail.com

Lee, Joo-Kyeong (Corresponding author)
Professor, Department of English Language and Literature
University of Seoul
163, Seoulsiripdae-ro, Dongdaemun-gu
Seoul, Korea
Tel: +82-2-6490-2519
Email: jookyeong@uos.ac.kr

* This work was supported by the 2025 Research Fund of the University of Seoul for Joo-Kyeong Lee.

ABSTRACT

Han, Ye-Yeong and Joo-Kyeong Lee. 2026. Measuring vowel contrast revisited. *Korean Journal of English Language and Linguistics* 26, 275-285.

This study revisits the measurement of vowel contrast by evaluating the Pillai score with a sample-size-adjusted threshold (P_{95}) as a categorical criterion. We examined the assumption that vowels are contrastive when the Pillai score exceeds the P_{95} threshold as derived from the Stanley and Sneller (2023)'s contrapositive. Six adjacent English vowel pairs (/i–ɪ/, /ɛ–æ/, /u–ʊ/, /ɑ–ɔ/, /ɑ–ʌ/, /ɔ–ʌ/) produced by 36 native speakers in two speech corpora were analyzed through MANOVA on normalized F1, F2, and duration. Pillai scores of six vowel pairs were calculated for each speaker and then converted to binary values (P_{bin}), coded as 1 if the score exceeded P_{95} and 0 otherwise. Results showed that /i–ɪ/ and /ɑ–ʌ/ consistently exceeded the threshold across *all* speakers, while the /u–ʊ/ pair exhibited the lowest rate of threshold achievement as found in 27 out of 36. This indicates that 27 native speakers did not produce these vowels contrastively, reflecting unstable phonemic status of the /u–ʊ/ pair. Logistic regression confirmed that all vowel pairs were significantly more likely to yield P_{bin}=1. We concluded that native speakers tend to produce vowel pairs contrastively, although some speakers exhibited merged productions for certain pairs. The application of the Pillai score with its P_{95} threshold to vowel contrast analysis provides a unified criterion for assessing vowel contrasts in English. This approach could be further applied in L2 contexts to eliminate the need for direct native-speaker comparison.

KEYWORDS

vowel contrast, MANOVA, Pillai, Threshold

1. Introduction

Vowel analysis has been a fundamental concern in acoustic phonetics since the early development of speech science, with researchers continuously seeking more precise and comprehensive methods to quantify vowel systems. Understanding and quantifying vowel contrasts is an elemental task in the fields of phonetics and phonology. Vowel contrasts—differences in vowel sounds that distinguish meaning—are often analyzed through acoustic measurements, particularly formant frequencies.

A variety of methodologies have been proposed for quantifying vowel contrasts. Among them, Euclidean distance—calculated between two vowel points in F1–F2 space—has been widely used for the analysis of comparison. Flege et al. (1997) analyzed the mean spectral differences between /i-ɪ/ and /ɛ-æ/ using Euclidean distance between two vowel points in each pair. They converted formant frequency values to the Bark scale and calculated vowel height and frontness/backness by subtracting these values (B1–B0 and B2–B1). Each non-native speaker's vowel production was compared to the native norm by measuring the Euclidean distance between their B1–B0 and B2–B1 values and the native group's average values. Chung (2016) also used Euclidean distance to compare vowel contrasts produced by Korean learners of English and native English speakers. The results demonstrated a significant group effect for all vowel contrasts—/i-ɪ/, /ɪ-ɛ/, /ɛ-æ/, /æ- ɑ/, /ɔ-ʌ/, /ʌ-ʊ/, and /ʊ-u/—except /ɑ-ɔ/.

Nycz and Hall-Lew (2014) highlighted several limitations associated with the use of Euclidean distance in vowel analysis. It only serves as a comparative measure, as the outcome is a continuous distance value. While this metric allows comparison of distances, it does not establish a threshold for determining whether a given distance reflects a contrastive distinction. In other words, it constrains interpretation regarding whether the contrast is meaningfully distinct. Moreover, they asserted that it cannot reflect the distributional feature of two vowels or quantify the degree of their overlap. Because of its mathematical nature, it relies solely on mean values, disregarding variability within each category.

These limitations have led researchers to adopt statistically robust methods such as ANOVA (Analysis of Variance) or linear regression for univariate analysis of individual acoustic dimensions, including F1, F2, and duration. Both ANOVA and linear regression are univariate statistical analyses, which compare mean values of a single dependent variable at a time. Therefore, if F1, F2, and duration are analyzed, three separate tests are required. Lee and Rhee (2019) examined how L2 vowel pronunciation relates to proficiency level by conducting three ANOVAs to analyze F1, F2, and duration separately. Park and Lee (2024) investigated the developmental progression of English vowel production among Korean learners by interpreting multiple separate linear regressions, each conducted for F1, F2, and duration. These methods are statistically robust, providing p-values that refer to whether the two groups are significantly different.

Still, these methods have several disadvantages. Statistically, they compare means without accounting for distributional characteristics. In addition, analyzing each parameter separately can complicate interpretation. Specifically, the question of whether two vowels are contrastive depends on how many acoustic parameters show significant differences. For instance, a vowel pair differing significantly in three parameters is presumably more contrastive than one differing in only two. Yet contrastiveness is not inherently gradient, but rather a categorical distinction. As such, ANOVA and linear regression are limited in addressing contrastiveness between two vowels in question. That is, they do not explain how many parameter differences are necessary to establish vowel distinctiveness. Furthermore, they do not consider the distributional patterns of acoustic parameters among vowels.

To overcome these issues, researchers have employed various multivariate approaches that assess the distributional overlap between vowel categories. The vowel overlap refers to the degree of the acoustic similarity between the vowel categories. In dialectology, the overlap measurement has been used in the studies about vowel mergers (Hay et al. 2006, Haynes and Taylor 2014, Morrison 2008, Wassink 2006). A high degree of overlap between two vowel categories indicates that a vowel merger is either ongoing or has already occurred (Kelley and Tucker 2020). Kelley and Tucker (2020) compared four types of vowel overlap measurements: SOAM (Wassink 2006), APP-based metric (Morrison 2008), VOACH (Haynes and Taylor 2014), and Pillai score (Hay et al. 2006). They assessed four measures using Monte Carlo simulations and concluded that the Pillai score performed best.

The Pillai score, also known as the Pillai-Bartlett trace, is derived as part of MANOVA and is used to test for significant multivariate effects across groups (Pillai 1955). While ANOVA is limited to a single dependent variable, MANOVA can evaluate multiple dependent variables simultaneously. MANOVA provides holistic statistical analysis by yielding a single p-value for multiple dependent variables. Therefore, it reduces ambiguities of interpreting multiple univariate tests. This makes MANOVA particularly suitable for the analysis of acoustic features, where dependent variables such as F1 and F2 should be considered jointly. Conventionally, researchers have referred to the Pillai-Bartlett trace simply as the Pillai score, or just Pillai (Havenhill 2024, Hay et al. 2006, Lee 2023, Li et al. 2023, Stanley and Sneller 2023). The Pillai score quantifies the degree of the overlap between the two classes, with values ranging from 0 to 1. Higher values indicate greater distinction, while lower values suggest more overlap (Hay et al. 2006).

Hay et al. (2006) introduced the Pillai score to quantify vowel overlap, followed by studies of vowel merger (Havenhill 2024, Lee 2023, Li et al. 2023). In these studies, the Pillai score was used primarily for comparative purposes, rather than for interpreting its absolute value, due to the absence of a definitive threshold indicating complete merger. Stanley and Sneller (2023) conducted Monte Carlo simulations using merged vowel pairs to derive the 95th percentile (P_{95}) of the Pillai score distribution. Through 96,000 simulations testing sample sizes from 5-100, they discovered that when both Pillai scores and sample sizes were log-transformed, the 95th percentile followed a straight line with intercept 1 and slope -1. This observation led to the derivation of the formula $P_{95} = e/m$, where e is Euler's number and m is the average sample size per vowel group. Since their study examined vowel mergers, they proposed that when vowels are merged, Pillai score is less than or equal to P_{95}, 95% of the time. In the context of vowel merger, P_{95} has been used to serve as a threshold value. Lee (2023) investigated the merger of mid vowels in Seoul Korean, employing the formula proposed by Stanley and Sneller (2023) to calculate P_{95} and interpret Pillai scores based on this cut-off. He concluded that the mid vowel merger remains incomplete, particularly among male speakers, who exhibited signs of the merger approximately one generation later.

2. Assumption

Building on Stanley and Sneller (2023)'s investigation of vowel merger, their study reports that when two vowels are merged, the corresponding Pillai score is less than or equal to P_{95}. In the present study, we take this empirical generalization to constitute a true proposition. Under standard principles of propositional logic, the contrapositive of a true conditional statement is also logically true. This relationship is summarized in (1).

(1) Stanley and Sneller (2023)'s proposition and its contrapositive

a. Proposition

If vowels are merged, then the Pillai score is less than or equal to P_{95}.

b. Contrapositive

If the Pillai score is *not* less than or equal to P_{95}, then the vowels are *not* merged.

A crucial issue raised by the contrapositive concerns the interpretation of '*not* merged'. Vowel merger is widely understood as a gradual, diachronic process, during which originally contrastive vowel categories may progressively lose their phonemic distinction. If vowels are not merged, this implies that they continue to function as separate, contrastive phonemes within the system. Accordingly, the negation of merged can be logically interpreted as contrastive. Adopting this interpretation, we formulate the central assumption of the current study in (2).

(2) Assumption of the current study

If the Pillai score is greater than P_{95}, then the vowels are contrastive.

This assumption allows any pair of vowels in a language to be evaluated quantitatively: vowels may be classified as merged or contrastive depending on whether their Pillai score falls below or exceeds the P_{95} threshold. English was selected as the test language because it has a relatively large vowel inventory (nine stressed monophthongs), which necessarily results in reduced articulatory and acoustic space between vowel categories. Previous research has argued that the vowels /ɔ/ and /ɑ/ have merged in certain varieties of American English (Clopper et al. 2005). In the present study, six vowel pairs in American English were constructed and examined to determine whether their Pillai scores exceed the corresponding P_{95} values. Specifically, the degree of non-overlapping distribution between the two vowels was assessed using acoustic dimensions such as F1, F2, and duration, and evaluated against the 95th percentile threshold. If the merger between /ɔ/ and /ɑ/ has progressed to a substantial extent in American English, their Pillai score is expected not to exceed P_{95} in the corpora analyzed in this study.

3. Method

We selected 6 adjacent pairs of English monophthongs for analysis: /i-ɪ/, /ɛ-æ/, /u-ʊ/, /ɑ-ɔ/, /ʌ-ɑ/, and /ʌ-ɔ/. These pairs have been considered challenging, as they are distributed closer in spectral and/or temporal dimensions. They were selected due to their proximity in the vowel space. Some neighboring vowels are contrasted in different combinations of multiple phonetic features. For example, Park and Lee (2024) found that native speakers differentiate the tense /i/ vowel and the lax /ɪ/ vowel which are expected to be both spectrally and temporally distinct based on all three features like tongue height (F1), tongue frontness (F2), and duration, whereas the phonemic contrast of the lax vowels /ɛ/ and /æ/ was maintained merely by tongue height (F1) and duration, without involving tongue frontness (F2).

This study utilized acoustic data from two corpora: the ALLSSTAR corpus and the Korean-English Intelligibility corpus (Bradlow n.d.). Only native speakers were included in this study. All speech consisted of read separate sentences or paragraphs. Additional information related to the corpora is available on the Northwestern University SpeechBox website (https://speechbox.linguistics.northwestern.edu). Recordings from 36 native speakers were segmented using Montreal Forced Aligner (McAuliffe et al. 2017) and analyzed with a PRAAT script to extract F1 and F2 at vowel midpoints and duration of the target vowel.[1] In this process, only stressed vowels were selected. Duration was normalized as it may vary depending on speech rate. To do so, we calculated articulation rate (syllables per millisecond) by dividing the number of syllables by the time of sound production. The calculation of duration × articulation rate was then used as a speech rate adjusted measure of duration. F1 and F2 formant values were normalized using Lobanov's (1971) z-score transformation method.

We first conducted vowel-level analyses by submitting the normalized data to MANOVA for each vowel pair across all speakers, with vowel category as the independent variable and F1, F2, and duration as dependent variables. This analysis yielded six Pillai scores, and a corresponding P_{95} value was assigned to each score using the formula *e/m*, where *e* denotes Euler's number and *m* corresponds to half of the sample size, following the procedure outlined by Stanley and Sneller (2023). Each Pillai score was then evaluated to determine whether it exceeded its corresponding P_{95} threshold.

To examine individual speaker variability in the production of vowel pairs, we then conducted speaker-level analyses by fitting separate MANOVAs for each speaker using the same model specification, resulting in 216 tests (36 speakers × 6 vowel pairs). Each vowel pair for a given speaker was assigned its own P_{95} value. For each instance, the Pillai score was converted into a binary variable (P_{bin}), coded as 1 if it exceeded the threshold (P_{95}), and 0 otherwise. For statistical testing, P_{bin} was conducted using logistic regression, an appropriate method for modeling binary outcome variables.

4. Results

4.1 Vowel-level

The combined data from all speakers are presented in Table 1. Sample size varied across vowel pairs, ranging from 2,241 for /u–ʊ/ to 7,292 for /ɛ–æ/, with corresponding P_{95} values ranging from 0.000746 for /ɛ–æ/ to 0.00243 for /u–ʊ/. Interestingly, the Pillai score of six pairs ranges from 0.0376 for /u-ʊ/ to 0.497 for /i-ɪ/. The P_{bin} represents whether the Pillai score exceeds its corresponding P_{95} value, where 1 indicates that the Pillai score was greater than the threshold, and 0 indicates that it was less than or equal to the threshold. The results showed that Pillai scores for all vowel pairs exceeded their corresponding P_{95} thresholds, yielding P_{bin} values of 1.

Table 1. Summary of Vowel-Level Analyses

Vowel Pair	Pillai Score	Sample Size	P_{95}	P_{bin}
/i-ɪ/	0.497	6,674	0.000815	1
/ɛ-æ/	0.0967	7,292	0.000746	1
/u-ʊ/	0.0376	2,241	0.00243	1
/ɑ-ɔ/	0.182	4,280	0.00127	1
/ɑ-ʌ/	0.334	5,531	0.000983	1
/ɔ-ʌ/	0.263	5,603	0.000970	1

[1] Given the size of the corpora, it was not feasible to manually inspect all data for potential errors arising from automatic proc essing. Instead, we examined the target vowels in the first three sentence recordings for each speaker, focusing on the accurac y of TextGrid alignments as well as F1 and F2 formant values and vowel durations. As no errors were detected in this subset, we proceeded with the subsequent analyses.

4.2 Speaker-level

Table 2 shows descriptive statistics for six vowel pairs. Among these, the /i-ɪ/ contrast exhibited the highest mean Pillai score (*M* = 0.546, *SD* = 0.127), suggesting the clearest distinction between the two vowel categories among the vowel contrast pairs. In contrast, the /ɛ-æ/ pair showed the lowest mean Pillai score (*M* = 0.148, *SD* = 0.0665). The /u-ʊ/ contrast showed the widest range (0.002 to 0.775) and the highest variability (*SD* = 0.228), reflecting considerable individual differences in vowel production. For the low vowel pairs, /ɑ-ɔ/ had a mean Pillai score of 0.301 (*SD* = 0.134, range = 0.019–0.532), reflecting a moderate level of contrast. The /ɑ-ʌ/ distinction was slightly stronger (*M* = 0.387, *SD* = 0.153, range = 0.071–0.623). Finally, the /ɔ-ʌ/ pair yielded a mean score of 0.438 (*SD* = 0.220, range = 0.041–0.699), marking it as one of the more distinct contrasts among the set.

Sample sizes also varied across vowel pairs. For /i-ɪ/, the mean sample size was 185.0 (*SD* = 67.1, range = 60–262), and for /ɛ-æ/, it was slightly higher at 203.0 (*SD* = 90.3, range = 56–276). In contrast, the /u-ʊ/ pair had the smallest sample size (*M* = 62.2, *SD* = 31.0, range = 12–90), reflecting a more limited dataset for this contrast. Among the low back vowels, /ɑ-ɔ/ had a mean sample size of 119.0 (*SD* = 39.9, range = 48–151), while /ɑ-ʌ/ showed a higher mean of 154.0 (*SD* = 41.2, range = 80–205). The /ɔ-ʌ/ pair was similar, with a mean of 156.0 (*SD* = 65.5, range = 43–219). Given that variability in sample size is a known factor affecting Pillai scores, P_{95} was calculated separately for each vowel pair within each speaker.

Table 2 also shows the number of speakers with a P_{bin} values of 1 and 0. For the /i-ɪ/ and /ɑ-ʌ/ pairs, all 36 speakers had a P_{bin} value of 1, which represents that every Pillai score exceeded the corresponding threshold. In the case of /ɛ-æ/, /ɑ-ɔ/, and /ɔ-ʌ/ pairs, the number of speakers with a P_{bin} value of 1 was 32, 34, and 35, respectively. In contrast, the /u-ʊ/ pair showed the lowest distinction, with 27 speakers exceeding the threshold. Overall, there were much more speakers who showed that $P_{bin} = 1$ for each pair.

Table 2. Summary of Speaker-Level Analyses

Vowel Pair	Mean Pillai Score (SD, Range)	Mean Sample Size (SD, Range)	$\mathbf{P_{bin} = 1}$	$\mathbf{P_{bin} = 0}$
/i-ɪ/	0.546 (0.127, 0.137–0.733)	185.0 (67.1, 60–262)	36	0
/ɛ-æ/	0.148 (0.0665, 0.022–0.256)	203.0 (90.3, 56–276)	32	4
/u-ʊ/	0.249 (0.228, 0.002–0.775)	62.2 (31.0, 12–90)	27	9
/ɑ-ɔ/	0.301 (0.134, 0.019–0.532)	119.0 (39.9, 48–151)	34	2
/ɑ-ʌ/	0.387 (0.153, 0.071–0.623)	154.0 (41.2, 80–205)	36	0
/ɔ-ʌ/	0.438 (0.220, 0.041–0.699)	156.0 (65.5, 43–219)	35	1

For statistical testing, logistic regression was conducted with vowels pairs as independent variables and with P_{bin} of 1 and 0 as dependent variables. Table 3 indicates that every vowel pair showed a statistically significant likelihood of $P_{bin} = 1$. For two pairs, /i-ɪ/ and /ɑ-ʌ/, all tokens were assigned to the same bin ($P_{bin} = 1$), indicating no variation in the binary outcome. Therefore, fitting a logistic regression model was unnecessary for these pairs. The outcome suggests that these contrasts were consistently produced above the threshold. Among the remaining pairs, /ɛ-æ/ reached significance with p = 0.0001, confirming that the contrast was distinguished despite its relatively low mean Pillai score. The /u-ʊ/ contrast also yielded significance (p = 0.0043), though the effect was weaker compared to other pairs. Both /ɑ-ɔ/ and /ɔ-ʌ/ were highly significant (p = 0.0001 and p = 0.0005, respectively), reflecting stable production of contrastiveness.

Table 3. The results of logistic regression on P_{bin} for each vowel pair

Vowel Pair	Estimate	*p*-value	Note
/i-ɪ/	25.60	NA	All tokens in the same bin ($P_{bin} = 1$)
/ɛ-æ/	2.08	0.0001*	
/u-ʊ/	1.10	0.0043*	
/ɑ-ɔ/	2.83	0.0001*	
/ɑ-ʌ/	25.60	NA	All tokens in the same bin ($P_{bin} = 1$)
/ɔ-ʌ/	3.56	0.0005*	

* Significant

Figure 1 presents the relationship between sample size and Pillai scores across vowel contrasts, supporting the use of binary classification based on P_{95}. In Figure 1, the x-axis represents sample size, and the y-axis shows the corresponding Pillai score. The shape of each point indicates the value of P_{bin}: filled circles represent $P_{bin} = 1$, empty circles represent $P_{bin} = 0$. Because two different corpora were used, the sample size exhibited two distinct distributions, and the threshold was adjusted accordingly based on sample size.

As shown in Figure 1(a), the /i-ɪ/ pair has all 36 filled circles, which indicates that all 36 speakers had a P_{bin} value of 1. In other words, their Pillai scores all exceeded the P_{95} values, indicating that the two vowels were produced as distinct categories by all speakers. Likewise, Figure 1(e) shows that all points for the /ɑ-ʌ/ pair are represented as filled circle, which means these vowels were produced distinctively by every speaker. As depicted in Figure 1(b), four speakers failed to exceed the threshold in the /ɛ-æ/ pair (empty circles). Figure 1(d) indicates that the /ɑ-ʌ/ pair have two empty tokens ($P_{bin} = 0$). In Figure 1(f), the /ɔ-ʌ/ pair exhibits only one speaker with an empty circle, whose Pillai score did not exceed its threshold. In case of the /u-ʊ/ pair, in Figure 1(c), a total of nine tokens are represented by empty circles, indicating that these nine speakers failed to exceed the P_{95} threshold. This may be related to the relatively small size of the /u-ʊ/ pair, which has the smallest mean sample size (62.2). This will be discussed in detail in Section 5. Discussion.

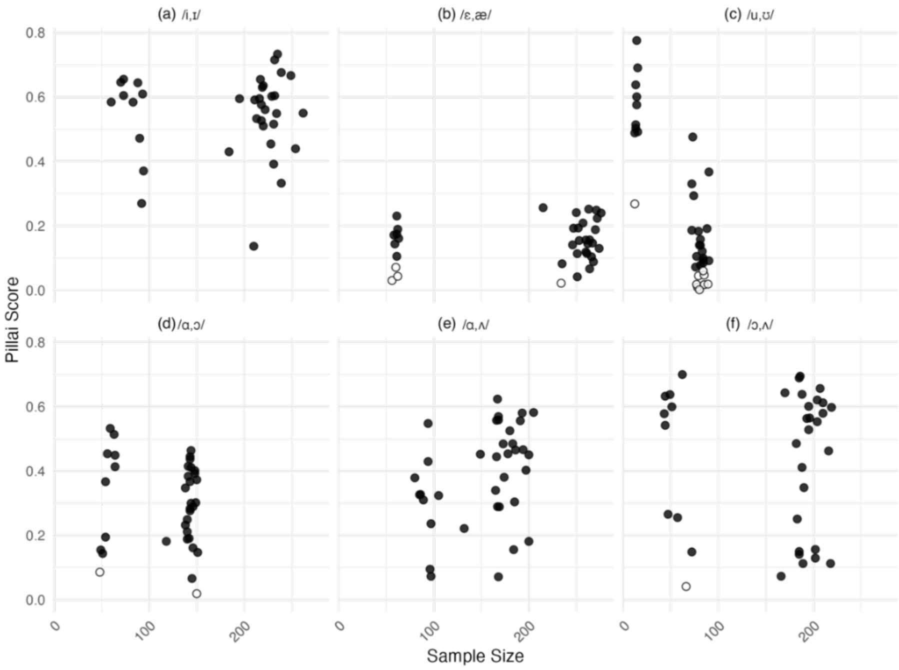

Figure 1. Relationship Between Sample Size and Pillai Score for Each Vowel Pair

5. Discussion

The current study examined the assumption that vowels are contrastive when the Pillai score exceeds the P_{95} threshold, using two native corpora. Following the proposal of Stanley and Sneller (2023), we adopted its contrapositive: if the Pillai score exceeds the P_{95} threshold, the production of a vowel pair is contrastive.

The vowel-level analyses revealed that, when data from all speakers were combined, Pillai scores for all six vowel pairs exceeded their corresponding P_{95} threshold ($P_{bin} = 1$). This indicates that native productions of these six adjacent vowel pairs were contrastive rather than merged. The sample-size–adjusted threshold (P_{95}) enabled meaningful comparisons across vowel pairs with differing sample sizes. The corpora used in this study contained unequal numbers of tokens for each vowel pair; for example, the /ɛ–æ/ pair comprised 7,292 tokens, whereas the /u–ʊ/ pair comprised 2,241 tokens. This variability is explicitly incorporated into the threshold equation, which yields different P_{95} values for different vowel pairs. Consequently, Pillai scores evaluated against the P_{95} threshold are appropriately scaled to datasets with varying sample sizes.

The logistic regression results from the speaker-level analyses showed that all six vowel pairs had a statistically significant likelihood of achieving $P_{bin} = 1$, indicating that more speakers produced the vowel pairs distinctively. Nevertheless, some speakers exhibited merged productions for certain vowel pairs. As shown in Table 2, while no speaker produced /i-ɪ/ and /ɑ-ʌ/ as merged, /ɛ-æ/, /u-ʊ/, /ɑ-ɔ/, and /ɔ-ʌ/ were produced as merged by 4, 9, 2, and 1 speakers, respectively. Interestingly, the /ɑ–ɔ/ pair was produced as merged by only two speakers. The /ɑ–ɔ/ merger has been reported in some regional varieties of American English, including New England, the Mid-Atlantic, the Midland, and Western regions (Clopper et al. 2005). Unfortunately, speaker-level regional background information was, however, unavailable in the corpora used in this work, and we were unable to further investigate why only two out of 36 speakers exhibited the merger of /ɑ–ɔ/ productions. One possible explanation is that these speakers may have originated from other regions than the areas reported above.

Notably, in Table 2, the /u-ʊ/ pair showed the lowest rate of threshold achievement (27 out of 36 speakers) despite having intermediate Pillai scores (M = 0.249). The /u–ʊ/ vowel pair exhibited the substantial individual variation of its mean Pillai score (SD = 0.228, range = 0.002–0.775). This indicates that the /u–ʊ/ contrast was less stable than the other vowel pairs within a speaker. Their phonemic status appeared to be less robust, which prompted further acoustic analyses. We conducted a linear regression analysis to identify which acoustic features – F1, F2, and/or duration – most strongly influenced their less stable contrast. The results revealed significant effects of duration ($p < .05$) and F1 ($p < .05$), whereas F2 did not reach significance ($p = .0514$). This indicates that the reduced stability of the /u–ʊ/ contrast is primarily driven by its reliance on temporal cues (duration) and vowel height (F1), whereas the contribution of vowel frontness/backness (F2) is relatively weak. In other words, speakers appeared to distinguish /u-ʊ/ more through duration and tongue height than through tongue backness and/or lip rounding.

This finding has a theoretical importance in the phonetic measurement of vowel contrasts. Unlike traditional methods such as Euclidean distance, which provides only continuous values rather than interpretable criteria, and ANOVA or linear regression, which require determining how many independent variables should be statistically significant in each pair of vowels, the binary values of P_{bin} (the Pillai score with its P_{bin} threshold) provide an integrated interpretation using multivariate acoustic features. It enables testing whether the categorical classification of vowel pairs is contrastive or merged. This suggests a significant advancement in vowel analysis methodology, providing researchers with not only comparative but also absolute measures of vowel contrast. Notably, Mairano et al. (2023) employed the Pillai score to evaluate L2 vowel production without direct comparison to native speaker data. They found that the Pillai score was a strong predictor of native listeners' ratings of both comprehensibility and nativelikeness. However, they did not account for the effect of the sample size. Given that size critically affects the Pillai score, incorporating a sample-size-adjusted binary classification (P_{bin}) would further enhance the utility of the Pillai score in L2 speech research contexts.

The establishment of reliable threshold for vowel contrast measurement may have significant implication for second language acquisition research. Previous studies have relied on separate univariate tests for multiple parameters, having a limitation in assessing whether L2 learners have achieved native-like vowel contrasts (Lee and Rhee 2019, Park and Lee 2024). The Pillai score with threshold can provide a unified criterion that could serve as a benchmark for assessing L2 vowel learning process. For example, researchers can calculate the Pillai score for L2 learners' productions of specific vowel pairs and then compute the corresponding threshold P_{95} to determine their P_{bin} value. If their P_{bin} equals 1, it would provide quantitative evidence that they have acquired the contrast. Conversely a P_{bin} value of 0 would indicate incomplete acquisition or ongoing difficulty with that contrast. This approach offers two important advantages. First, as mentioned in Mairano et al. (2023), it does not require direct comparison with native speakers. Instead, the P_{bin} value itself provides an internal benchmark for evaluating whether a given vowel pair is contrastive. Second, the use of P_{bin} values would allow contrastiveness to be represented categorically (0 or 1), rather than as a continuous numeric score. This categorical distinction makes it easier to interpret whether learners have acquired a vowel contrast, avoiding the ambiguity that can arise from relying solely on raw Pillai scores.

One limitation of the present study is the lack of detailed speaker background information in the corpora, particularly regarding regional origin. As a result, potential regional differences in vowel contrastiveness or merger patterns could not be examined. In addition, vowel contrast in this study was assessed exclusively based on speakers' productions. While production-based measures provide important evidence of phonetic contrast, they do not necessarily reflect how such contrasts are perceived by listeners. Factors such as listeners' perception and sociolinguistic background were not considered. Future research integrating perceptual data and social factors would provide a more comprehensive understanding of vowel contrast and merger.

6. Conclusion

This study analyzes the Pillai score with a sample-size-adjusted threshold (P_{95}) for assessing vowel contrasts. By applying Stanley and Sneller (2023)'s framework of vowel merger in a contrapositive manner, we established a quantitative criterion (P_{bin}) to determine whether vowel pairs are contrastive or not. The results show that this approach effectively captures distinctions among American English monophthongs in native speech, although some merged productions were detected. It offers researchers a holistic criterion for determining vowel distinctiveness that solves key limitations of traditional methods. Importantly, this method would allow the evaluation of vowel contrasts in non-native speech without requiring direct comparison to native speakers. It also overcomes interpretive challenges associated with continuous acoustic measures by providing a clear, categorical indicator of contrastiveness.

References

Bradlow, A. R. n.d. ALLSSTAR: Archive of L1 and L2 Scripted and Spontaneous Transcripts and Recordings. Available online at https://speechbox.linguistics.northwestern.edu/allsstar

Bradlow, A. R. n.d. Scripted speech corpora – Korean-English intelligibility. Available online at https://speechbox.linguistics.northwestern.edu/scripted

Chung, Y-H. 2016. Categorial discrimination in the perception and production of English vowels by Korean EFL learners: A case study. *Studies in English Language & Literature* 42(2), 263-282.

Clopper, C. G., D. B. Pisoni and K. de Jong. 2005. Acoustic characteristics of the vowel systems of six regional varieties of American English. *The Journal of the Acoustical Society of America* 118(3), 1661-1676.

Flege, J. E., O-S. Bohn and S. Jang. 1997. Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics* 25(4), 437-470.

Havenhill, J. 2024. Audiovisual enhancement of vowel contrast: Production and perception of the cot-caught contrast in Chicago. *Laboratory Phonology* 15(1).

Hay, J., P. Warren and K. Drager. 2006. Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics* 34(4), 458-484.

Haynes, E. F. and M. Taylor. 2014. An assessment of acoustic contrast between long and short vowels using convex hulls. *The Journal of the Acoustical Society of America* 136(2), 883-891.

Kelley, M. C. and B. V. Tucker. 2020. A comparison of four vowel overlap measures. *The Journal of the Acoustical Society of America* 147(1), 137-145.

Lee, S-H. and S-C. Rhee. 2019. The relationship between vowel production and proficiency levels in L2 English produced by Korean EFL learners. *Phonetics and Speech Sciences* 11(2), 1-13.

Lee, Y-C. 2023. Revisiting the merger of mid front vowels in Seoul Korean. *Studies in Phonetics, Phonology and Morphology* 29(2), 193-209.

Li, P., J. E. Flege, C. D. Martin and N. Kartushina. 2023. Speech stability across time: Evidence from Norwegian vowels in spontaneous speech production. In *Proceedings of the 20th International Congress of Phonetic Sciences* (ICPhS 2023), 1177-1181.

Lobanov, B. M. 1971. Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America* 49(2B), 606-608.

Mairano, P., F. Santiago and L. C. Roa. 2023. Can L2 pronunciation be evaluated without reference to a native model? Pillai scores for the intrinsic evaluation of L2 vowels. *Languages* 8(4), 280.

McAuliffe, M., M. Socolof, S. Mihuc, M. Wagner and M. Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of Interspeech 2017*, 498-502.

Morrison, G. S. 2008. Comment on "A geometric representation of spectral and temporal vowel features: Quantification of vowel overlap in three linguistic varieties" [J. Acoust. Soc. Am. 119, 2334–2350 (2006)]. *The Journal of the Acoustical Society of America* 123(1), 37-40.

Nycz, J. and L. Hall-Lew. 2014. Best practices in measuring vowel merger. In *Proceedings of Meetings on Acoustics* 20(1), 060008.

Park, M-S. and J-K. Lee. 2024. Korean ESL Learners' production of English vowel contrasts: Developmental variations in L2 sound learning. *Korean Journal of English Language and Linguistics* 24, 1318-1332.

Pillai, K. C. S. 1955. Some new test criteria in multivariate analysis. *The Annals of Mathematical Statistics* 26(1), 117-121.

Stanley, J. A. and B. Sneller. 2023. Sample size matters in calculating Pillai scores. *The Journal of the Acoustical Society of America* 153(1), 54-67.

Wassink, A. B. 2006. A geometric representation of spectral and temporal vowel features: Quantification of vowel overlap in three linguistic varieties. *The Journal of the Acoustical Society of America* 119(4), 2334-2350.

Examples in: English
Applicable Languages: English
Applicable Level: Tertiary