Konin Language Studies
Faculty of Humanities and Social Sciences, University of Applied Sciences in Konin, Poland
KSJ 11 (2). 2023. 141-165
http://ksj.konin.edu.pl
doi: 10.30438/ksj.2023.11.2.3

Beyond temporal dominance: Reassessing spectral and temporal cues in second language speech using multilingual corpus data

Jahurul Islam ✉
University of British Columbia, Vancouver, Canada
https://orcid.org/0000-0001-9147-7610
jahurul.islam@ubc.ca

Sayeed Anwar
Rajshahi University of Engineering & Technology, Bangladesh
https://orcid.org/0000-0003-0913-2766
anwar@hum.ruet.ac.bd

Abdulla Al Masum
Green University of Bangladesh, Purbachal American City, Bangladesh
https://orcid.org/0009-0004-1632-8294
masum.gulc@green.edu.bd

Abstract

The relative importance of temporal and spectral cues in non-native speech sound production and perception has been explored in numerous studies (Bohn, 1995; Ćavar et al., 2022; Gao et al., 2020; Yazawa et al., 2023). The prevailing view in the existing literature mostly suggests that second language (L2) speakers tend to prioritize temporal cues over spectral cues (Perwitasari, 2019; Yuan & Archibald, 2022); however, some recent studies (e.g., Ćavar et al., 2022) have reported mixed patterns. To enhance our comprehension of the utilization of temporal and spectral cues in L2 speech, the present study employs a larger-scale corpus analysis of L2 English with the L2-ARCTIC corpus (Zhao et al., 2018), incorporating data from six L1 backgrounds. This study compares the duration, F1, and F2 of tense and lax vowels, evaluating the relative significance of these cues in tense-lax distinctions through linear mixed-effects models and random forest classification models. Surprisingly, the results do not align with the prevalent notion that speakers prioritize temporal cues over spectral cues. Instead, the speakers exhibited a greater preference for spectral cues, in contrast to previous research. These findings offer novel insights into the interplay between temporal and spectral cues in non-native speech production and perception.

Keywords: spectral cues; temporal cues; L2 vowels; tense-lax contrast; English

1. Introduction

The understanding of how non-native speakers acquire and perceive speech sounds in a second language (L2) has been a topic of significant interest in linguistics and language acquisition research and a fundamental aspect of this process is the relative importance of temporal and spectral cues in acquiring second language vowel contrasts. Extensive experimental studies (Gao et al., 2020; Yazawa et al., 2023) have consistently reported that L2 learners tend to place more emphasis on temporal properties than on spectral properties when learning non-native vowel sounds. Such reports are, however, mostly based on perceptual or small-scale production data of the L2 learning process, and recent evidence from recent studies (Ćavar et al., 2022; Islam et al., 2023) indicates that the preference for temporal cues to spectral cues may not be generalizable to all speech modality and to all languages. To further investigate the question, the current study aims to evaluate the debates around the relative importance of temporal and spectral cues in L2 speech with a larger speech corpus containing speech from six different languages.

The existing literature reveals a significant gap in the comprehensive understanding of the relative importance and roles of temporal and spectral cues in the audition and production of L2 speech. Although both cues are fundamental to L2 speech learning, their precise contributions remain less understood. Research has often approached these cues in extremely controlled laboratory settings (Gao et al., 2020; Yazawa et al., 2023) or with participants from a single first language (L1) background (Chen, 2006; Rojczyk, 2010). This fragmented approach has resulted in a limited understanding of how these cues function beyond controlled experiments and in more naturalistic settings, where sounds are more complex and dynamic. Furthermore, there is a lack of consensus on the relative effectiveness of these cues across different tasks and language backgrounds, complicating the ability to generalize findings across contexts. This limitation highlights a critical need for studies that utilize more holistic and nuanced methods to explore the synergy between temporal and spectral cues in L2 speech. The current study addresses these gaps by employing a novel approach that combines analysis techniques from frequentist statistics and machine learning, facilitating a triangulation of findings through machine learning methods. Unlike previous experimental studies, this study adopts a corpus approach to include a large collection of near-natural speech from individuals with six different L1 backgrounds. This approach allows us to test whether conclusions from experimental studies on the roles of temporal and spectral cues in L2 speech are generalizable beyond tightly controlled settings and whether findings from studies focusing on a single language can be extended to other, unexplored languages, indicating a universal pattern. By addressing these critical gaps, the study aims to advance our theoretical understanding and practical implications of the roles of temporal and spectral cues in L2 speech learning.

2. Literature review

Numerous experimental studies (Bohn & Flege, 1990; Ćavar et al., 2022; Gao et al., 2020) have delved into the relative importance of temporal and spectral cues in non-native speech sound production and perception. Consistent findings from existing research assert that, during non-native vowel acquisition, speakers tend to prioritize temporal cues over spectral cues. Several investigations (e.g., Ćavar et al., 2022; Chen, 2006; Fox & Maeda, 1999; Gao et al., 2020; Podlipský et al., 2019; Rojczyk, 2010; Yazawa et al., 2023; Yuan & Archibald, 2022) have been conducted to examine whether L2 speakers exhibit a language-independent preference for durational features over spectral features. The remainder of this section will provide a comprehensive exploration of this prevailing trend across various studies and elucidate its implications for our understanding of non-native sound acquisition processes.

Research into the significance of cues in L2 vowel acquisition can be traced back to the 1990s. Bohn (1995) conducted a perception study to examine how individuals from Spanish and Mandarin language backgrounds, characterized by distinct L1 vowel systems, perceived English tense/lax vowel contrasts (/i/ vs. /ɪ/ – as observed in words like *beat* and *bit*) while acquiring English as their second language (L2). Worth noting, Spanish and Mandarin both exhibit a solitary phonemic category (/i/) for vowel sounds, whereas English encompasses two (/i/ and /ɪ/) within a shared acoustic range. The study employed synthesized auditory stimuli, presenting participants with tokens spanning the phonetic spectrum from *beat* to *bit*. Participants were assigned the task of identifying these tokens. The findings indicated that native English speakers predominantly attended to spectral distinctions, with limited consideration for the temporal aspect during the identification task. Conversely, participants originating from Spanish and Mandarin linguistic backgrounds heavily relied upon duration cues to discriminate between tense and lax vowels.

Fox and Maeda (1999) investigated Japanese speakers to explore whether Japanese speakers could differentiate between the two high front vowels (/i/ and /ɪ/) of American English in both production and perception perspectives. Perception data were collected from 12 native Japanese speakers who completed a two-choice forced identification task. The production study was completed by 7 speakers where they read out lists of words containing the target vowels in specific environments. Results of the perception study revealed that the participants were consistently successful in distinguishing between the two vowel sounds based on vowel duration whereas there was variation in vowel quality. Similar results were reported from the production study where participants produced consistent durational differences between /i/ and /ɪ/. Thus, the authors concluded that Japanese speakers tend to prioritize temporal cues over spectral cues in their L2 English vowels.

Rojczyk (2010) undertook a study involving Polish speakers who were learners of English as a second language (L2), comprising both production (43 participants) and perception (17 participants) data. The focal point of the study was primarily centered around investigating the influence of duration on the production and perception of the vowel /æ/ in relation to its neighboring vowels: /e/ and /ʌ/. The findings revealed that these learners displayed a predilection for durational cues over spectral cues when distinguishing among the English vowels /æ/, /e/, and /ʌ/. Analyses of durational measurements obtained from the production study highlighted that speakers harnessed durational differences among these vowels, especially in cases where their spectral qualities were similar. The outcomes of the production study were substantiated by the results of the perception study, in which listeners exhibited a marked inclination to associate stimuli with lengthier vowel durations as representing /æ/, as opposed to /ʌ/.

Chen's (2006) study centered on Mandarin learners of English as their L2, exploring the spectral and temporal attributes in English tense and lax vowels using both production and perception data. The study concentrated on six American English vowels: /i, ɪ, æ, ɛ, u, ʊ/. These vowels were recorded in the context of /hVd/, embedded within a carrier phrase. The study cohort comprised 40 native Mandarin speakers for the production part. In the perception component, the authors ran tasks reminiscent of an AX task where the stimuli contained unmanipulated vowels sourced from the production data gathered earlier in the study. Based on the analysis of duration and formant data, the authors concluded that Mandarin speakers exhibited a heightened reliance on temporal features when contrasted with their native English-speaking counterparts. Also, tense-lax contrasts were reported to be less distinctive among Mandarin speakers, in both production and perception, compared to the reference group.

Unlike previous experimental studies, Yazawa et al. (2023) conducted a corpus analysis with Japanese learners of L2 English, aiming to comprehensively assess the utilization of spectral and temporal cues across all English vowel categories. The study incorporated production data collected from 102 native Japanese speakers, who read English passages aloud. The research explored various models of second language acquisition and reported a mixed array of results. The study highlighted that temporal implementation's effectiveness hinged on the successful execution of prosodic elements like stress patterns and phrase-final lengthening. Proficiency levels in the L2 were found to influence cue preferences; individuals with less native-like L2 proficiency showed a pronounced reliance on their L1 categories in their speech production, underlining the influence of these categories in shaping spectral implementation.

While the majority of studies exploring the comparative preference and impact of temporal and spectral cues in L2 speech production have predominantly focused on English as the L2, Gao et al. (2020) undertook a study involving Mandarin speakers who were learning German as their L2. The investigation collected production data from 30 Mandarin Chinese speakers who were acquiring German as their second language. These participants produced both the tense vowels and their corresponding lax vowel counterparts (/a:, o:, e:, i:, u:, y:, ø:/ and /a, o, ɤ, ɪ, u, y/, respectively) in the German language. The placement of these vowels was between the sounds /d/ and /t/, forming isolated monosyllabic nonce words within the context of /dVt/. Through the analysis of acoustic features, the study revealed that Mandarin speakers exhibited a heightened reliance on duration to differentiate between the tense and lax vowel pairs. In contrast, German speakers (the reference group) leaned more significantly on distinctions in formant frequencies.

As the above review shows, the majority of previous studies have reported that L2 learners tend to heavily rely on the durational properties of vowels rather than spectral properties. However, some counter-evidence has also been reported in recent research. For instance, Ćavar et al. (2022) examined the distinction between tense and lax vowels in L2 English, focusing on Polish and Croatian learners (33 Polish and 14 Croatian participants), using AXB perception experiments. The study manipulated both duration and quality in two pairs of English tense-lax vowels (/ɪ/--/i/ and /ʊ/ and /u/), resulting in a total of 50 stimuli (25 for the front pair and 25 for the back pair). The objective was to discern the dominant cues (temporal or spectral) on which the learners relied. The research unveiled intriguing patterns in their categorization strategies, with Croatian participants primarily depending on duration cues, while Polish participants leaned towards vowel quality, which encompasses spectral cues.

In addition, Islam et al. (2023) conducted a study investigating the significance of durational and spectral cues in the second language (L2) tense and lax vowel distinctions of English, as produced by Bangla speakers from Bangladesh. The study aimed to test previous assertions suggesting that speakers predominantly rely on durational cues rather than spectral cues when distinguishing L2 tense and lax vowel pairs. For this study, citation-style production data were collected from 16 native Bangla speakers, all of whom were undergraduate students. The data collection involved a shadowing task, where participants listened to a carefully curated list of real English words presented randomly and immediately repeated each word upon hearing it. The results revealed a noteworthy departure from the general patterns observed in other languages. Bangla speakers, in contrast to prior claims, did not prioritize durational cues for differentiating English tense-lax vowel pairs. Instead, they exhibited a preference for spectral cues over durational cues in this specific L2 context.

The preceding studies, as highlighted above, suffer from several limitations. Primarily, many of these studies have depended on experimental data derived from controlled environments, thereby raising concerns about their ecological validity. The deliberate control inherent in experimental designs leads to reduced data variability and limited dataset sizes, which in turn constrains the examination of how variables might behave in more natural, real-world contexts. To achieve deeper understanding of the generalizability of findings to diverse scenarios, it becomes imperative to validate the predictions of experimental studies on a larger scale, encompassing natural variability. Notably, the necessity of larger datasets, comprising at least 200 tokens per category, to effectively address the breadth of variability has been emphasized in previous research (Whalen, 2023; Whalen & Chen, 2019). Furthermore, earlier studies have predominantly relied on data sourced from single languages, often constrained to languages such as Japanese, Mandarin, and German. For a more comprehensive understanding, it is essential to draw insights from a wider array of languages. Additionally, the methodological disparities among various studies complicate cross-language result comparisons. To facilitate enhanced comparability, it is imperative to assess these inquiries through a multilingual dataset while adhering to a consistent methodology. Also, most of the previous studies picked only a few vowel categories; this can provide insights about specific vowel contrasts, but for a broader understanding, we need to include more vowel categories in a single study.

In our pursuit to expand the discourse on the significance of temporal and spectral cues in L2 speech, our study contributes to the existing literature by embracing a corpus-based approach. While Yazawa et al. (2023) also adopted a corpus methodology, their study's focus diverged somewhat. Our chosen approach offered a wealth of data, enabling the careful selection of 78,203 tokens from a pool of over 150,000 raw vowel tokens. We also chose to include all 4 tense-lax vowel pairs in English (/i-ɪ/, /e-ɛ/, /o-ɔ/, and /u-ʊ/), thereby broadening our scope. Moreover, this our encompasses balanced data sourced from speakers representing six distinct languages, many of which (such as Arabic, Hindi, Vietnamese, and Korean) have not been extensively studied before in the context of these research questions. By employing comparable data across these six languages, we were able to employ a uniform set of analytical methods, thereby enhancing the potential for comparability and providing a broader perspective. Further elaboration on the dataset particulars is presented in the subsequent sections. To summarize, here is the primary research question of this study:

> *Do L2 speakers universally put more emphasis on temporal cues over spectral cues in producing L2 vowel contrasts, irrespective of their L1 background?*

3. Methods

3.1. Data

The data in this study were obtained from L2-ARCTIC corpus[1] (Zhao et al., 2018) of L2 English. The corpus contains L2 speech from 6 different L1 backgrounds (with balance in gender, language, and tokens per vowel). The languages included: Arabic, Chinese, Hindi, Korean, Spanish, and Vietnamese. There were 4 speakers (2 males and 2 females) for each of the languages, providing a total of 24 speakers. According to the corpus documentation[2], participants read out sentence prompts from the CMU ARCTIC set[3] while being audio-recorded at 44.1 kHz sampling frequency. The corpus also included time-aligned phone-level annotations as Praat (Boersma & Weenink, 2023) TextGrids that were generated via force-alignment with Montreal Forced Aligner v1.0.0 (McAuliffe et al., 2017).

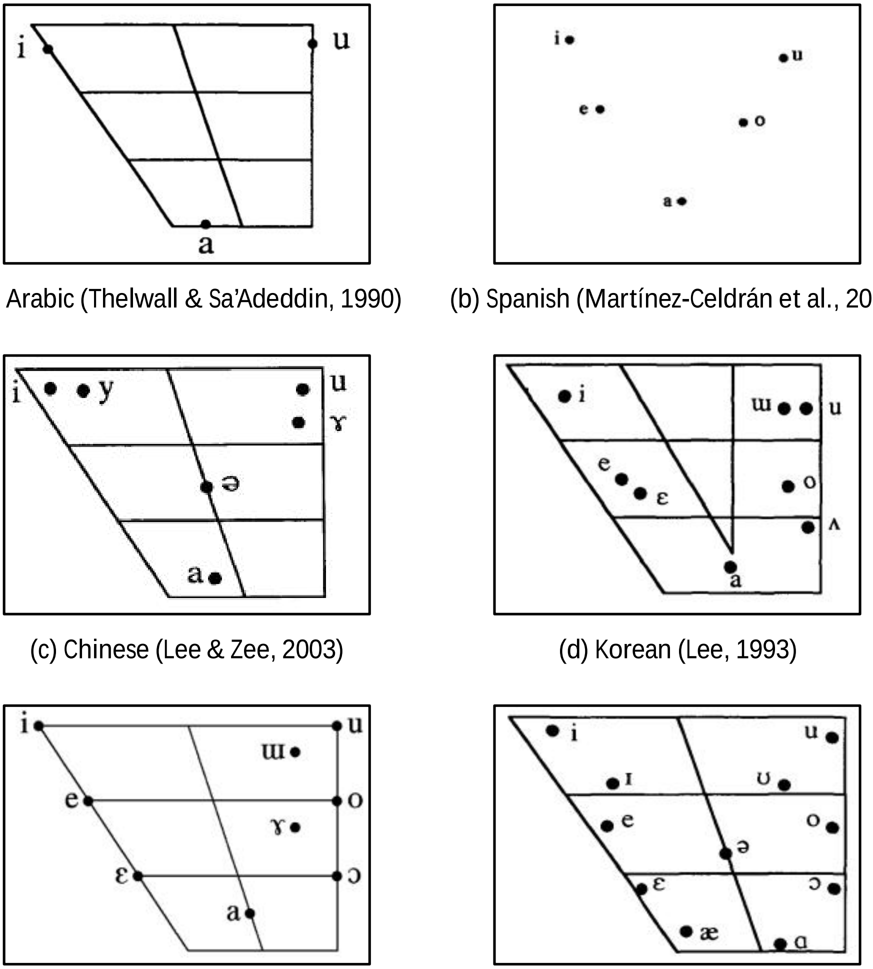
It may be useful to have a quick background on vowel inventories of the six languages included in this study; Figure 1 presents the basic monophthong inventories in the languages involved. As the figure shows, Arabic is the language with the least number of vowel categories in the inventory with only three peripheral vowels, followed by Spanish which has 5 basic vowels. Chinese can also be considered a 5-vowel system (excluding schwa since it is often not realized as

[1] http://festvox.org/cmu_arctic/cmuarctic.data

[2] Downloaded from https://psi.engr.tamu.edu/l2-arctic-corpus/, Jun 10, 2023.

[3] https://psi.engr.tamu.edu/l2-arctic-corpus/

a full vowel in many languages); however, unlike Spanish which utilizes height differences, Chinese employ rounding distinctions among high vowels to achieve the contrasts. Korean has a very interesting system with only one tense-lax contrast (/e/ vs /ɛ/) and only one rounding contrast (/u/ vs /ɯ/) in an 8-vowel inventory. Korean also has duration contrasts for all these vowels; thus, even though the inventory is not very large, Korean employs a lot of featural contrasts. Vietnamese has a vowel inventory which is very similar to Korean. Finally, the Hindi vowel system is very similar to the 11-vowel English inventory with vowels with the exact same phonological contrasts. Thus, it would be interesting to see how the findings of the study are generalizable to the speakers of all these diverse vowel systems.



(a) Arabic (Thelwall & Sa'Adeddin, 1990)

(b) Spanish (Martínez-Celdrán et al., 2003)

(c) Chinese (Lee & Zee, 2003)

(d) Korean (Lee, 1993)

(e) Vietnamese (Kirby, 2011)

(f) Hindi (Ohala, 1994)

Figure 1 Schematic vowel inventories in the six languages included in this study

3.2. Variables and measures

We used FastTrack (Barreda, 2021) which is a plug-in software for Praat (Boersma & Weenink, 2023) to extract vowel formants and duration measurements for the tense and lax vowels. FastTrack was chosen over the traditional methods of formant measurement because FastTrack generates multiple formant objects for each sound file and then automatically picks the best option with assistance from regression analyses of the estimated trajectories. While the traditional method is prone to providing erroneous formant measurement, which often requires manual intervention, FastTrack is robust against such measurement errors. For each vowel, the variables we measured included duration (in ms), F1 (in Hz), F2 (in Hz), stress, and the speaker's sex. Measurements for duration, F1, and F2 were taken at vowel midpoints for monophthongs and at 33% into the vowel for EY and OW (which typically have diphthongal quality).

The list of vowels measured included EY, IH, IY, EH, UW, UH, OW, OY, AO, ER, AH, AA, AE, AY, and AW (the vowel symbols here follow the ARPAbet convention), and the number of total vowels measured was 216,444. Out of the 15 measured categories, diphthongs (except EY and OW) and rhotic vowels were excluded; this left 10 vowels: EY, IH, IY, EH, UW, UH, OW, AO, AA, AE before we ran vowel normalization. In addition, to avoid co-articulatory effects, vowel tokens adjacent to nasals, rhotics and glides were also excluded. Finally, duration, F1 and F2 were normalized within speakers using the Lobanov method of vowel normalization (Thomas, 2017). The normalization was done using the full vowel system (all 10 vowels); after that, only the vowels participating in tense-lax contrasts were retained (i.e., AA and AE were excluded). This left us a final total of 78,203 vowel tokens. Table 1 presents the token counts for each vowel category across all six languages.

Table 1 Token counts by vowels and languages

Language	IY	IH	EY	EH	OW	AO	UW	UH
Arabic	2991	3908	1237	1225	923	727	922	370
Chinese	2995	4268	1326	1385	990	696	1053	393
Hindi	2888	4210	1329	1346	980	743	1055	407
Korean	3155	4421	1309	1403	971	884	920	393
Spanish	3414	4064	1291	1317	975	715	891	390
Vietnamese	3213	4294	1319	1421	985	694	996	401

3.3. Why we chose tense/lax vowels for this study:

Tense and lax vowel contrast presents a compelling opportunity to assess the significance of durational and spectral cues in second language acquisition. This contrast involves the utilization of both duration and formant values, primarily F1 and

F2, across various English dialects. Focusing on tense-lax distinctions allows us to explore the interplay between durational and spectral cues, with a particular emphasis on their relative importance in the context of second language acquisition. Furthermore, by incorporating data from multiple languages, we can investigate whether the preference for these cues is influenced by language-specific constraints or whether it is governed by language-independent general cognition. More specifically, it would be interesting to see whether having a larger set of contrasts in the vowel space facilitates learning contrasts in L2.

3.4. Data analysis

The primary objective of this study was to analyze the relative importance or effect of durational and spectral features in producing a vowel distinction based on tense-lax contrast. To achieve this goal, the data were analyzed from a number of angles. First, the variance in duration, F1 and F2 were compared statistically to determine if the duration has a lower variance compared to the spectral features. Second, the relative contributions of duration, F1, and F2 to the tense-lax contrast were assessed via Linear Mixed-Effects models (as regression tasks). Finally, the relative importance of the same features for tense lax contrasts was also investigated via random forest models (as classification tasks). Additionally, the success of distinguishing tense-lax contrasts was investigated via MANOVA models. All statistical analysis in this study was performed using R (R Core Team, 2023); however, Python (Van-Rossum & De Boer, 1991) was used for a Random Forest classification task.

4. Results

This section presents the empirical findings of the study, structured into four distinct parts. The initial part offers a descriptive analysis of the variables measured, focusing on the distinctions between tense and lax vowels. The second part details the outcomes of our analysis using linear mixed-effects models. Subsequently, the third part reports on the results derived from a random forest classification task. The final part explores how speakers differentiated between tense and lax vowels within a two-dimensional space.

4.1. Descriptive exploration

Figure 2 presents the distributions of the duration values in all tense and lax vowel categories across the six languages. The x-axis shows all the vowel categories while the y-axis represents the normalized (z-score transformed) duration values; shading is used to distinguish between tense and lax vowels. As the figure shows, tense vowels tend to be longer than their lax counterparts, in general, as indicated by the higher median, Q1, and Q3 values; and this trend is common to speakers of all six languages. An exception to this trend was the OW-AO pair where the lax vowel appears to be longer than or of similar duration to the tense one. This data provides an indication that L2 speakers aim for a longer duration in tense vowels, which consequently is consistent with the proposal that L2 speakers do attend to durational cues.

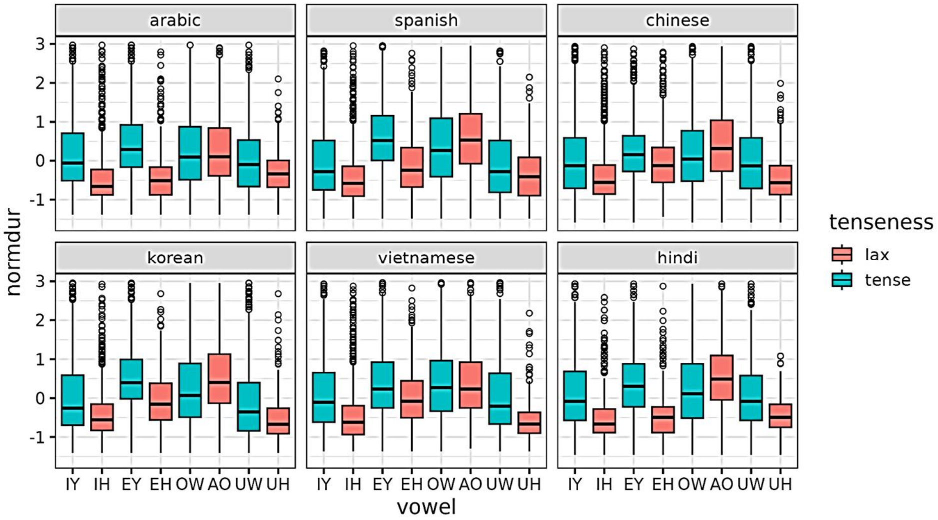


Figure 2 Distribution of duration values (normalized) in tense and lax vowels

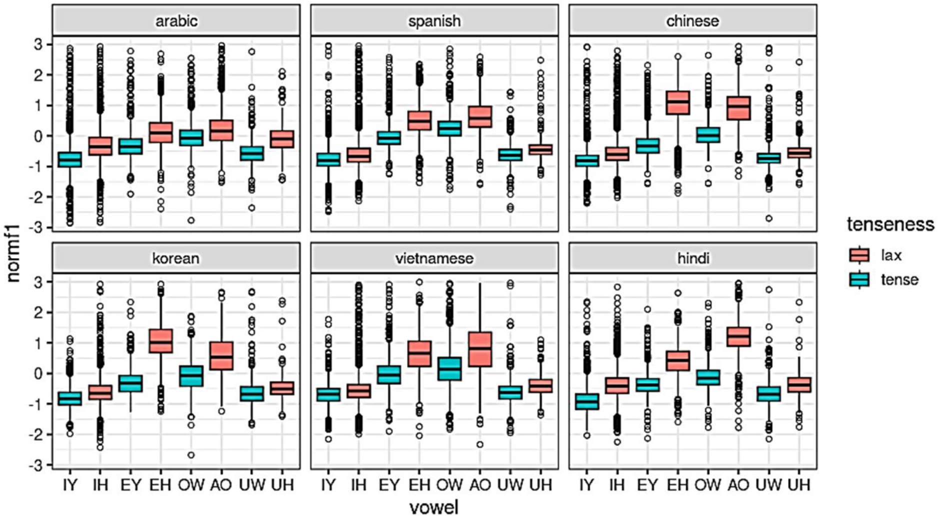


Figure 3 Distribution of F1 values (normalized) in tense and lax vowels

Figure 3 presents the F1 values in tense and lax vowels. As before, the y-axis represents normalized (z-score transformed) F1 values. As per the general characteristics of English vowels, tense vowels are expected to have lower F1 values (since they are higher in the vowel space) compared to their lax counterparts. As the figure demonstrates, that trend is quite consistently confirmed here; all the tense vowels are seen to have higher median, Q1 and Q3.

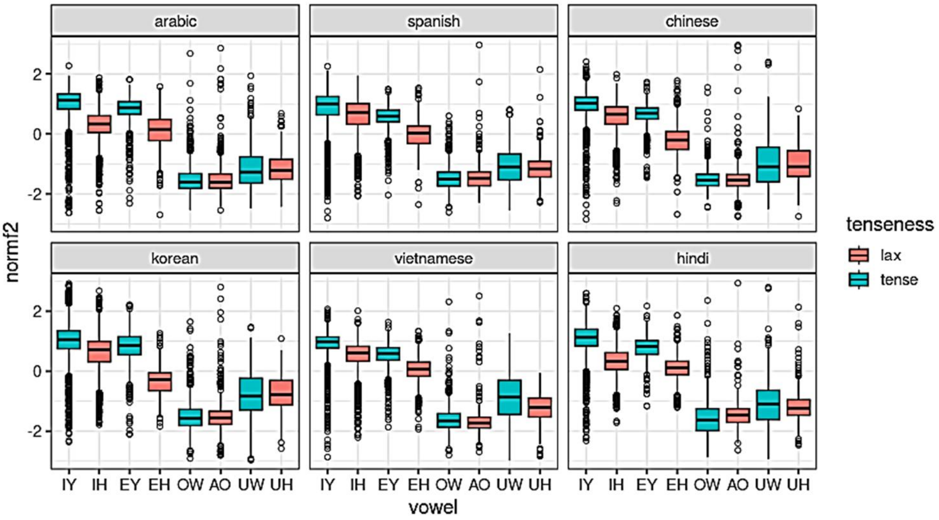


Figure 4 Distribution of F2 values (normalized) in tense and lax vowels

Finally, Figure 4 presents the data for F2. As per the general expectations, front tense vowels are expected to have higher F2 values (since they are more peripheral or fronter in the vowel space). For high-back vowels, UW is supposed to have a lower F2 than UH; for mid-back vowels, the expectations are a bit mixed since the two vowels (OW and AO) can often have similar backness, and the difference is primarily conveyed via F1. As Figure 4 shows, front tense vowels had higher medians compared to their lax counterparts. The trends in the back vowels are mixed, as expected since there are no clear distinctions between the tense and lax vowels.

Results presented in Figures 2, 3, and 4 confirm that L2 learners use all of the three major cues (duration, F1, and F2) that are useful to distinguish tense and lax vowels. It is very interesting to find speakers of all six languages utilizing all three cues; this result is surprising because only one (Hindi) out of the six languages had all 8 tense-lax contrasts that are phonologically analogous to English. Two (Korean and Vietnamese) had only partial contrasts while the remaining three (Arabic, Spanish, and Chinese) had no tense-lax contrasts at all in their vowel inventory. This provides an indication of the existence of a language-independent mechanism that enables and promotes attention to both durational and spectral cues in L2 acquisition.

Since all speakers demonstrated the use of both temporal and spectral cues in vowel productions, a deeper exploration to determine the importance of these cues is still warranted. In the remainder of this paper, the question of relative importance has been investigated from three different angles/methods: (1) comparison of variance, (2) linear mixed-effects models, and random forest machine learning classification model. More details of this are provided in the relevant sections below.

4.2. Comparison of variability in duration, F1 and F2

Figure 5 shows the standard deviation of duration, F1 and F2, separately for each vowel category across all six languages. The x-axis shows all four tense vowels and their lax counterparts. The y-axis represents the standard deviation (calculated from normalized values) as a measure of variability in duration, F1 and F2 which are differentiated in different degrees of shading. Data from different languages are separately presented in different panels. The underlying assumption for comparing the variance in durational and spectral measurements here is that if L2 speakers put more emphasis on durational measurements than on spectral measurements, the distribution of durational measurements will be tighter (i.e., variance will be lower in duration values than in F1 or F2).

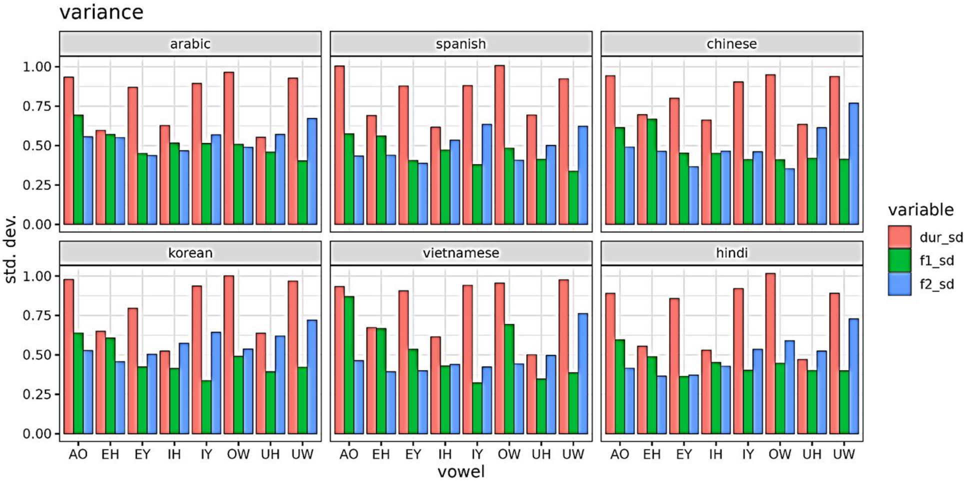


Figure 5 Variance (standard deviation) in duration, F1, and F2 in vowels across languages

As Figure 5 shows, *SD* values for duration are consistently higher than those for F1 and F2; the pattern is generalizable to most vowels and to all languages. In general, tense vowels have more variability (higher *SD* values) in duration, compared to their lax counterparts (except vowel AO); however, F1 and F2's variability is not much affected by a tense/lax distinction. Trends seen in this graph are completely inconsistent with the prediction that duration would have less variability compared to F1 and F2. Therefore, the data provides general evidence against duration being the preferable cue to L2 speakers.

To verify whether the differences between duration and F1 or F2 observed in Figure 5 are statistically significant, a series of Brown-Forsythes tests were performed (for all the vowel categories within each language) on two comparisons: duration vs. F1 and duration vs. F2. Brown-Forsythe's test is a statistical method used to compare the variances of two or more groups. It is a better alternative to the more commonly used F-test or Levene's test for comparing variances where the independence of samples assumption is not satisfied. Since the data in this study involve multiple measurements per speaker (aka repeated measures), Brown-Forsythe's test was chosen for its robustness against non-normal and non-independent data, such as clustered or repeated measures. The test was implemented using the *bf.test()* function from the R package *onesampletests* (Dag et al., 2023).

Table 2 presents Brown-Forsythes test results; for the sake of brevity, the table shows select comparisons (only the non-significant ones) out of the 96 pairwise comparisons run in total. For all other comparisons, the p-values were below .05. As observed, only 3 out of the 96 (3.12%) comparisons were non-significant; these results, combined with the trends seen in Figure 5, confirm that duration indeed did not have a lower variance than F1 or F2.

Table 2 Non-significant Brown-Forsythes Test results for the comparisons of duration vs. F1 and duration vs. F2. (3 out of 96 are reported here; all the remaining 93 statistically significant at an alpha level of .05 (87 of which were highly significant at an alpha level or *p* < .001))

Language	Vowel	Test variables	df1	df2	*F*	*p*
arabic	AO	normdur - normf1	1	1340.2	0.0	.914
chinese	UH	normdur - normf1	1	680.7	1.2	.281
korean	UH	normdur - normf1	1	652.9	0.4	.530

4.3. Mixed-effects models

Separate linear mixed-effects models were fitted within each language using the R package *lmerTest* (Kuznetsova et al., 2017). As fixed effects, each of these models included Duration, F1, F2, Stress, and Vowel-group. Of these, Duration, F1, and F2 were numeric variables; they were also scaled as z-scores. The Stress variable had three levels (1, 2, and 0) to indicate primary, secondary, and no-stress. The Vowel-group variable had four levels – high-front, mid-front, high-back, and mid-back – each having a tense vowel and its lax counterpart. The model also included random intercepts for speakers. The dependent variable in each of these models was the tenseness of the vowel (tense vs. lax). To enable a regression task, the levels of this variable were recoded as numerical values such that 1 is tense and 0 is lax.

The coefficients table from each fitted model was extracted using the summary() function in R. The effects of duration, F1, and F2 on the dependent variable were found to be statistically significant (*p* < .001) in each of the models (Table 3 presents more details). After confirming the significant effects of the three independent variables in concern, the beta-coefficients for each of them in the within-language models were compared to determine the relative differences in the magnitude of the effect they each had on the dependent variable. Figure 6 presents the beta-coefficient values for duration, F1, and F2 extracted from models fitted with individual languages. The x-axis represents the variables duration, F1, and F2 (all three were scaled as z-scores) while the y-axis represents the absolute beta-coefficient values (to avoid plotting bars in the negative directions). Since the model included z-score-normalized values for duration, F1 and F2, it enables the beta-coefficient values to be comparable between them (for these three variables), thereby enabling the comparability of the magnitude of their effect (i.e., effect size) on the dependent variable relative to each other.

Table 3 Coefficients from linear mixed-effects models from six languages

L1	Predictor	Estimate	Std. error	df	*t*	*p*
arabic	normdur	0.13	0.004	12293.85	29.21	< .001
arabic	normf1	-0.24	0.007	12294.96	-34.44	< .001
arabic	normf2	0.26	0.006	12294.07	40.31	< .001
chinese	normdur	0.08	0.004	13098.00	17.99	< .001
chinese	normf1	-0.31	0.007	13098.00	-46.63	< .001
chinese	normf2	0.20	0.007	13098.00	27.41	< .001
hindi	normdur	0.11	0.004	12949.40	27.56	< .001
hindi	normf1	-0.40	0.006	12949.79	-69.23	< .001
hindi	normf2	0.26	0.006	12949.90	44.11	< .001
korean	normdur	0.10	0.004	13445.65	23.51	< .001
korean	normf1	-0.33	0.007	13446.33	-48.02	< .001
korean	normf2	0.16	0.006	13446.60	25.94	< .001
spanish	normdur	0.11	0.007	13049.00	23.08	< .001
spanish	normf1	-0.26	0.009	13049.00	-30.67	< .001
spanish	normf2	0.15	0.008	13049.00	19.44	< .001
vietnamese	normdur	0.12	0.005	13313.95	26.97	< .001
vietnamese	normf1	-0.21	0.007	13313.21	-28.93	< .001
vietnamese	normf2	0.27	0.008	13309.57	34.35	< .001

As Figure 6 shows, duration had a consistently smaller effect on the dependent variable (tense- vs. lax-ness of a vowel) compared to F1 and F2. In each of the six languages, duration ranked lower than F1 and F2 in terms of the effect size. While the difference between the magnitude of the duration effect and that of the F2 effect is relatively smaller in Spanish and Korean, the F2 effect is still higher. Overall, the Figures show that spectral properties have a larger effect than duration properties on the tenseness of a vowel; and this finding is generalizable to all six languages studied here.

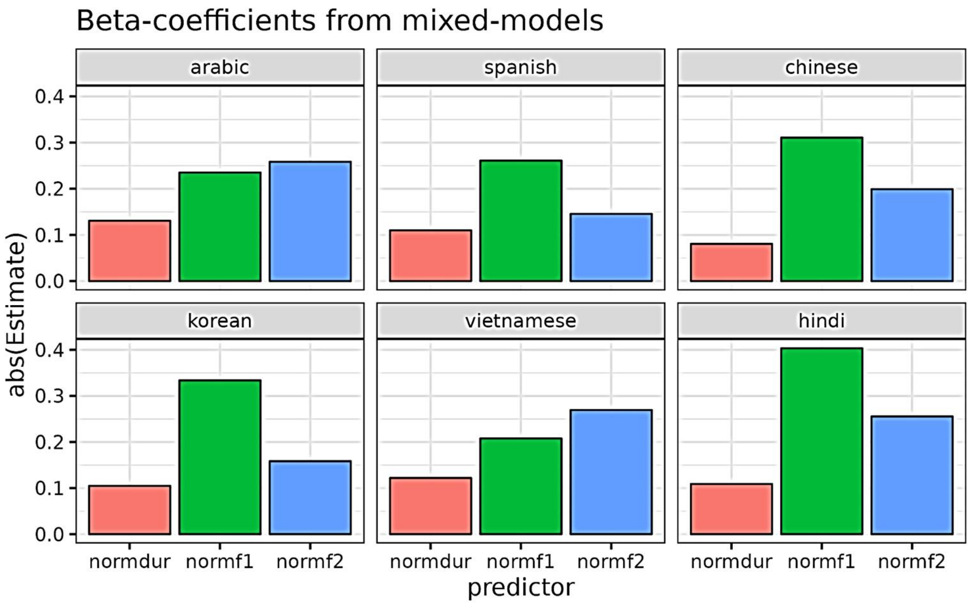


Figure 6 Comparison of beta-coefficients for duration, F1, and F2

4.4. Random forest models

The second approach to determine the relative importance of duration, F1 and F2 on the tenseness of a vowel involved fitting random forest classification models (Shaik & Srinivasan, 2019) and extracting feature-importance data from the fitted models within each language. Random Forest is a powerful ensemble learning algorithm widely used for classification tasks. It operates by creating multiple decision trees during training and combining their predictions to make more accurate and robust classifications. Each tree is built on a random subset of the training data and features, reducing overfitting and improving generalization. During classification, the algorithm aggregates the predictions from individual trees through a voting mechanism, where the most frequent class becomes the final prediction. Additionally, random forest allows the extraction of ranked feature-importance scores, enabling practitioners to identify the most influential features driving the classification performance. This feature makes it valuable for gaining insights into complex data relationships.

The dataset used for random forest modeling is the same as the one used for mixed-effects modeling above. Separate models were fit for each language. The target label was the tenseness variable (tense vs. lax) while the features used to train the models included duration, F1, F2, stress, and vowel group. As reported above, duration, F1, and F2 were numeric variables and were z-score scaled before feeding them into the models. Stress (3 levels: 1, 2, or 0) and vowel group (4 levels: high_front, mid_front, high_back, mid_back) were categorical variables and were converted into dummy variables.

Random forest classification models were implemented using *sklearn* Python library (Pedregosa et al., 2011). Data for each model were randomly split into train and test subsets using *train_test_split* function from sklearn where 80% of the records were used for training the model and the remaining 20% were used for testing the performance of the models. Candidates for the best set of hyperparameters were identified using sklearn's *GridSearchCV*, a technique for hyperparameter tuning. Various combinations of hyperparameters were systematically explored, and the configuration that yielded the highest accuracy on the dataset was determined. The final model was then trained using these best hyperparameters, ensuring optimal performance for the classification task. Accuracy scores were calculated using *accuracy_score()* function. Feature-importance scores were obtained via *model.feature_importances_* method.

Figure 7 reports the feature-importance scores for vowel duration, F1, and F2 within each language. The x-axis, again, represents the features (except the dummy variables) used to train the random forest models while the y-axis the feature-importance score for each of these features. The overall accuracy scores for each language are provided in the top-left corner of the corresponding panels.

As Figure 7 shows, duration was consistently ranked the lowest among the three features of concern and the trend is generalizable to all the six languages being studied. That is, duration was determined to be of lower importance than F1 and F2 for determining the tenseness of a vowel, and the difference between the feature-importance score for the duration and the two spectral features was consistently large. This provides a strong indication that the degree of association between duration and tenseness was considerably smaller than the degree of association between F1/F2 and tenseness. Thus, the evidence here indicates that L2 speakers did not rank duration over F1 and F2 when making a tense or lax distinction; rather, they put considerably greater emphasis on the spectral measures over the durational measures.

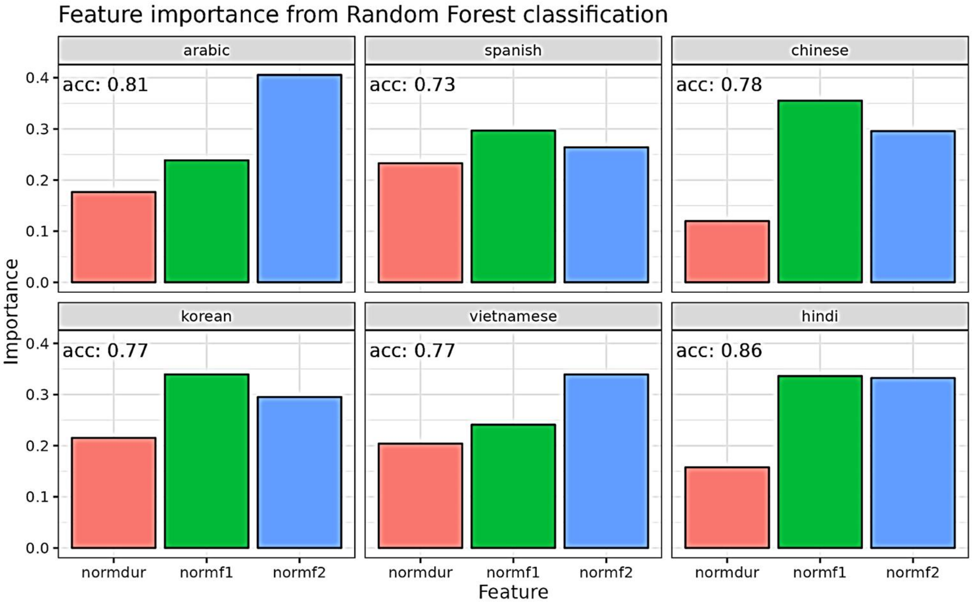


Figure 7 Feature-importance scores for duration, F1, and F2

4.5. Distinction between tense and lax vowels on F1xF2 plane

It is, however, important to note that the confirmation that all three cues (duration, F1, and F2) are actively used by all L2 speakers (at least for the six languages studied here) does not entail that speakers of all these languages were able to successfully distinguish the all the tense and lax vowel pairs; the task of producing the distinctions is a more complex one where at least a two-dimensional space (e.g., the F1 x F2 plane) is involved. To verify how successful speakers of different languages were in producing the tense/lax distinctions in vowel quality, Pillai-Bartlett scores (Pillai trace) were calculated from F1xF2 bivariate distributions via fitting MANOVA models for each tense-lax pair across all languages.

The Pillai score, also referred to as the Pillai-Bartlett trace, is a statistical metric derived from the output of a MANOVA model (more details can be found in Hall-Lew (2010)). MANOVA is a form of ANOVA that considers variations in multiple dependent variables simultaneously, such as both F1 and F2. The Pillai statistic value is bound between 0 and 1; a higher value indicates a greater distinction between the two distributions concerning these dependent variables. The use of Pillai statistic in vowel studies (primarily mergers) was introduced by Hay et al. (2006) and, since then, it has been used in many studies (e.g, Islam & Ahmed, 2020; Kennedy, 2006; Nycz & Hall-Lew, 2013; Wong & Hall-Lew, 2014).

Figure 8 presents the Pillai-Bartlett trace values for the degree of overlap between the tense vowels and their lax counterparts for each language. The x-axis represents the vowel pairs. The y-axis represents the Pillai-Bartlett trace value; a value approaching 0 (zero) indicates more overlap between the two vowel categories being compared while a value approaching 1 indicates less overlap (therefore, more distinct) between the two vowel categories. Different degrees of shading are used for the ease of comparisons between vowel pairs across languages (shown in panels). Pillai trace values were calculated separately for each speaker (each language had 4 speakers) and then the 4 values within each vowel pair were averaged to obtain a single Pillai trace value for each pair.

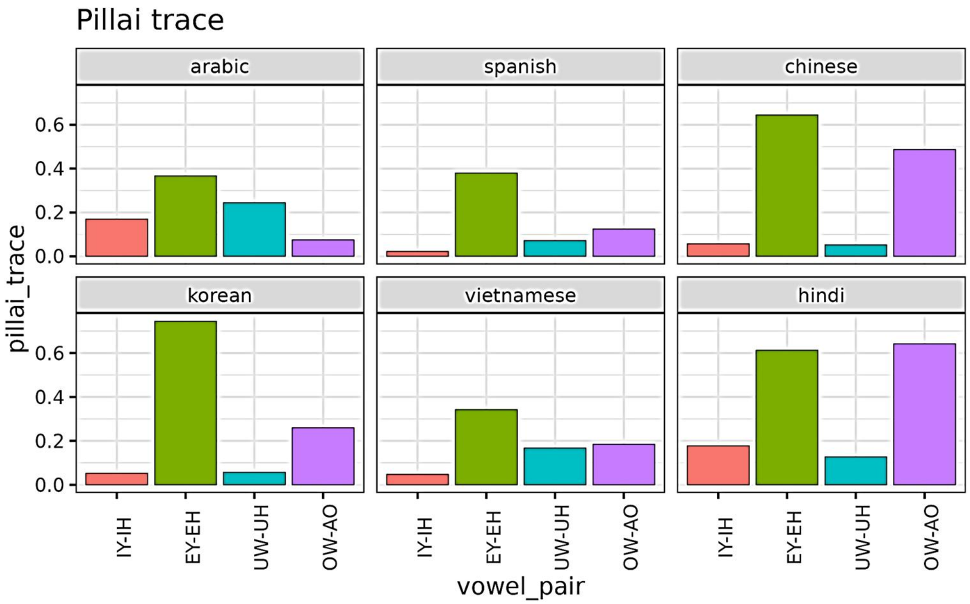


Figure 8 Pillai scores for the tense/lax distinctions in vowel pairs

As Figure 8 shows, speakers were not always successful in distinguishing between tense and lax vowels on a two-dimensional F1xF2 plane and the success varied considerably based on vowel pairs in concern. Speakers were most successful in distinguishing between EY and EH (mid-front) vowels and it is consistent across languages. This is followed by the mid-back pair OW-AO and this result is interesting because OW was not longer than AO in terms of duration (see Figure 2). Distinctions between the high vowels are relatively worse; for most languages, this might have been due to the non-existence of high tense-lax contrasts. The exception in Hindi (which does have high tense-lax pairs) could be due to the discrepancies between phonetic implementations of those vowels in English vs. Hindi. Korean speakers' high success in distinguishing EY and EH can be attributed to the existence of a front-mid tense-lax vowel contrast in the vowel inventory. But there can also be some language-independent factors here since we can see this category is the best in all languages.

5. Discussion

This study delved into the relative importance of temporal and spectral acoustic cues in L2 vowel production across six different languages, utilizing a large-scale corpus. Statistical techniques such as the Brown-Forsythe test for variance comparison, linear mixed-effects models to assess predictor effect sizes, and the random forest classification algorithm for relative featural importance were employed to evaluate the significance of vowel duration, F1, and F2 cues.

Results from this study challenge the prevailing assumption that L2 speakers prioritize temporal cues over spectral cues in the distinction of tense and lax vowels. Contrary to previous studies such as those by Fox and Maeda (1999), Gao et al. (2020), and Rojczyk (2010), the current investigation revealed a consistent and compelling trend: L2 speakers consistently accorded greater significance to spectral cues, specifically F1 and F2, in their vowel productions, outweighing the importance of durational cues. Moreover, a remarkable consistency was observed across all six languages under scrutiny. This uniform emphasis on spectral cues remained prevalent regardless of the speakers' diverse linguistic backgrounds. Notably, this convergence of patterns across different languages was validated through the application of three distinct analytical methodologies – variance comparisons, linear mixed-effects models, and random forest classification models. This confluence of evidence points towards a strikingly robust and language-independent propensity among L2 speakers to prioritize spectral cues over temporal cues when navigating the complexities of vowel distinctions.

The consistent prioritization of spectral cues over duration cues among speakers, as observed in this study, challenges prevailing expectations and offers valuable insights into mechanisms governing cue integration. This deviation from conventional findings in existing literature may be less unexpected than initially presumed, as it suggests that vowel perception and production are predominantly anchored in the analysis of spectral characteristics, particularly formant frequencies. This reliance on spectral attributes remains robust across languages, regardless of whether the native vowel inventory encompasses temporal differences. This observation aligns with linguistic typology, where vowel distinctions tend to be founded on spectral disparities before temporal variations come into play. Notably, languages lack instances of vowel inventories exclusively structured around temporal contrasts, such as inventories featuring solely /e/, /e:/, and /e::/, while inventories exclusively characterized by spectral differences, devoid of temporal distinctions, are commonly found, as exemplified in languages like Arabic and Spanish. This pattern highlights the pivotal role of spectral cues in shaping vowel distinctions and suggests a prevailing, language-independent tendency influencing vowel perception and acquisition.

As a potential explanation for L2 speakers' consistent maintenance of length distinctions between tense and lax vowels as effectively as F1 and F2 differences in our data, we posit that this phenomenon arises from speakers' active effort to achieve the gestural targets for those vowels. For instance, tense vowels, which are more peripheral in the vowel space and involve a stiffer tongue root, require greater effort and attention to achieve the gestural target, leading to increased production time compared to their lax counterparts.

The influence of cross-linguistic factors and the role of phonetic inventory also are important considerations in the interpretation of the findings. The consistent utilization of both spectral and durational cues by L2 speakers across diverse language backgrounds challenges the notion of cue selection as a language-specific phenomenon. Rather, the outcomes suggest the existence of a mechanism that facilitates higher attention to spectral than durational features during L2 acquisition, potentially indicating a universal cognitive predisposition. The assertion that L2 speakers prioritized spectral cues, irrespective of their native language's vowel inventory, underscores the significance of formant frequencies in vowel perception and production. This observation aligns with cross-linguistic and typological tendencies where languages tend to rely on spectral contrasts to distinguish vowels before introducing durational distinctions.

As the study revealed, speakers had varying degrees of success in distinguishing the tense and lax vowels in the four pairs. The consistently higher success in distinguishing between /EY/ and /EH/ in their productions compared to vowel pairs may stem from distinct acoustic and articulatory properties inherent to these specific vowels. Being a mid-vowel contrast in the front region of the oral cavity, /EY/ and /EH/ tend to exhibit more pronounced differences in their formant patterns (both F1 and F2) and associated articulatory gestures, contributing to their clearer acoustic differentiation. This heightened articulatory precision and attention to spectral cues during production could underlie the observed reliable distinction. The variable consistently in the discriminability of the vowel pairs could be attributed to potential cross-linguistic similarities in articulatory strategies, and a thorough investigation into these factors and their cross-linguistic generalizability remains to be seen in future investigations.

While our study provides some valuable insights, it has some limitations too. The focus on six languages is still limited in scope considering the number of living languages, and it may limit generalizability, and the investigation's scope is specific to tense and lax vowel distinctions. We also could not consider the proficiency level of L2 speakers. Future research could expand to include more languages, diverse phonetic features, and individual proficiency differences. Longitudinal studies could trace the evolution of cue integration strategies over time. Addressing these limitations and pursuing new avenues could advance our understanding of L2 vowel acquisition processes and cognitive mechanisms.

6. Conclusion

In conclusion, this study delves into the nuanced dynamics of L2 vowel acquisition, shedding light on the interplay between spectral and durational cues. The findings challenge traditional assumptions by revealing that L2 speakers consistently prioritize spectral cues over durational cues in their vowel productions, irrespective of their native language's phonetic inventory. This points toward a possible universal cognitive mechanism that integrates these cues during L2 acquisition. The study underscores the significance of cross-linguistic influences and phonetic inventory in shaping L2 cue weighting, offering insights into the complexities of L2 vowel acquisition processes. By advancing our theoretical understanding of these phenomena, this research contributes to the broader discourse on second language acquisition and highlights the need for further exploration of the intricate mechanisms underlying speech sound perception and production in non-native contexts.

References

Barreda, S. (2021). Fast track: Fast (nearly) automatic formant-tracking using Praat. *Linguistics Vanguard, 7*(1), 1-10.

Boersma, P., & Weenink, D. (2023). *Praat: Doing phonetics by computer* [Computer program]. Version 6.3.14. http://www.praat.org/

Bohn, O. S. (1995). Cross-language speech perception in adults: First language transfer doesn't tell it all. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 279-304). York Press.

Bohn, O. S., & Flege, J. E. (1990). Interlingual identification and the role of foreign language experience in L2 vowel perception. *Applied Psycholinguistics*, *11*(3), 303-328.

Ćavar, M. E., Rudman, E. M., & Oštarić. A. (2022). Temporal versus spectral cues in L2 perception of vowels: A study with Polish and Croatian learners of English. *Journal of Slavic Linguistics, 30*(1), 85-107.

Chen, Y. (2006). Production of tense-lax contrast by Mandarin speakers of English. *Folia Phoniatrica et Logopaedica*, *58*(4), 240-249.

Dag, O., Dolgun, A., Konar, N., Weerahandi, S., & Ananda, M. (2023). onewaytests: One-way tests in independent groups designs. R package version 2.7. https://CRAN.R-project.org/package=onewaytests

Fox, M. M., & Maeda, K. (1999). Perception and production of American English tense and lax vowels by Japanese speakers. *University of Pennsylvania Working Papers in Linguistics*, *6*(1), 299-314.

Gao, Y., Ding, H., & Birkholz, P. (2020). An acoustic comparison of German tense and lax vowels produced by German native speakers and Mandarin Chinese learners. *The Journal of the Acoustical Society of America, 148*(1), EL112-EL118.

Hall-Lew, L. (2010). Improved representation of variance in measures of vowel merger. In *Proceedings of meetings on acoustics, 9*(1), 1-10. AIP Publishing.

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

Islam, M. J., & Ahmed, I. (2020). Mid-front and back vowel mergers in Mymensingh Bangla: An acoustic investigation. *Linguistics Journal*, *14*(1), 206-232.

Islam, M. J., Masum, A., & Anwar, M. S. (2023). The role of temporal and spectral cues in non-native speech production: Bangla speakers' L2 English tense and lax vowels. *Crossings: A Journal of English Studies*, *14*, 130-153.

Kennedy, M., (2006). *Variation in the pronunciation of English by New Zealand school children* (Unpublished master's thesis). Victoria University of Wellington.

Kirby, J. (2011). Vietnamese (Hanoi Vietnamese). *Journal of the International Phonetic Association, 41*(3), 381-392. https://doi.org/10.1017/S0025100311000181

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software, 82*, 1-26.

Lee, H. (1993). Korean. *Journal of the International Phonetic Association, 23*(1), 28-31. https://doi.org/10.1017/S0025100300004758

Lee, W., & Zee, E. (2003). Standard Chinese (Beijing). *Journal of the International Phonetic Association, 33*(1), 109-112. https://doi.org/10.1017/S0025100303001208

Martínez-Celdrán, E., Fernández-Planas, A., & Carrera-Sabaté, J. (2003). Castilian Spanish. *Journal of the International Phonetic Association, 33*(2), 255-259. https://doi.org/10.1017/S0025100303001373

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using Kaldi. In *Interspeech* (Vol. 2017, pp. 498-502).

Nycz, J., & Hall-Lew, L. (2013). Best practices in measuring vowel merger. In *Proceedings of meetings on acoustics, 20*(1). AIP Publishing.

Ohala, M. (1994). Hindi. *Journal of the International Phonetic Association, 24*(1), 35-38. https://doi.org/10.1017/S0025100300004990

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Perwitasari, A. (2019). *English vowels produced by Javanese and Sundanese speakers* (Unpublished doctoral thesis), Leiden University.

Podlipský, V. J., Chládková, K., & Šimáčková, Š. (2019). Spectrum as a perceptual cue to vowel length in Czech, a quantity language. *The Journal of the Acoustical Society of America, 146*(4), EL352-EL357.

R Core Team (2023). *R: A language and environment for statistical computing. v.4.1.2.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rojczyk, A. (2010). Overreliance on duration in nonnative vowel production and perception: The within lax vowel category contrast. *Achievements and perspectives in SLA of speech: New Sounds*, *2*, 239-249.

Shaik, A. B., & Srinivasan, S. (2019). A brief survey on random forest ensembles in classification model. In *Proceedings of International Conference on Innovative Computing and Communications (ICICC) 2018*, 2, 253-260. Springer Singapore.

Thelwall, R., & Sa'Adeddin, M. A. (1990). Arabic. *Journal of the International Phonetic Association*, *20*(2), 37-39.

Thomas, E. (2017). *Sociophonetics: An introduction*. Bloomsbury Publishing.

Van Rossum, G., & De Boer, J. (1991). Interactively testing remote servers using the Python programming language. *CWI Quarterly, 4*(4), 283-303.

Whalen, D. (2023). Challenges of analyzing variability in speech from linguistic and motor control perspectives. In *Proceedings of the HISPhonCog 2023: Hanyang International Symposium on Phonetics & Cognitive Sciences of Language*. Hanyang University, Seoul, Korea, 68-69.

Whalen, D., & Chen, W. (2019). Variability and central tendencies in speech production. *Frontiers in Communication. 4*(49), 1-9.

Wong, A. W. M., & Hall-Lew, L. (2014). Regional variability and ethnic identity: Chinese Americans in New York City and San Francisco. *Language & Communication*, *35*, 27-42.

Yazawa, K., Konishi, T., Whang, J., Escudero, P., & Kondo, M. (2023). Spectral and temporal implementation of Japanese speakers' English vowel categories: A corpus-based study. *Laboratory Phonology*, *14*(1) 1-33.

Yuan, Q., & Archibald, J. (2022). Modified input training and cue reweighting in second language vowel perception. *Frontiers in Educational Research, 5*(6), 65-75.

Zhao, G., Sonsaat, S., Silpachai, A., Lucic, I., Chukharev-Hudilainen, E., Levis, J., & Gutierrez-Osuna, R. (2018). L2-ARCTIC: A non-native English speech corpus. In *Interspeech*. September. pp. 2783-2787. https://psi.engr.tamu.edu/l2-arctic-corpus/.