Korean Journal of English Language and Linguistics, **Vol 25, March 2025, pp. 289-310**
DOI: 10.15738/kjell.25..202503.289

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

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

http://journal.kasell.or.kr

Dynamic Cues in Vowel Classification: A Discriminant Analysis of Conversational Speech Corpus*

Hyun Jin Hwangbo (Pukyong National University)

Received: January 2, 2025
Revised: February 10, 2025
Accepted: March 8, 2025

Hwangbo, Hyun Jin
Assistant Professor, Division of English Language and Literature
Pukyong National University
Email: hjhwangbo@pknu.ac.kr

* This work was supported by a Research Grant of Pukyong National University (2023).

ABSTRACT

Hwangbo, Hyun Jin. 2025. Dynamic cues in vowel classification: A discriminant analysis of conversational speech corpus. *Korean Journal of English Language and Linguistics* 25, 289-310.

This paper asks whether the vowel inherent spectral change (VISC) or the dynamic cues of vowels is an essential feature for vowel classification in natural speech. To answer this question, vowels from the Buckeye Corpus of conversational speech were trained and tested for three models on vowel classification with quadratic discriminant analysis, a machine learning technique. Three models were evaluated: the steady-state model, the one-point model, and two trajectory models, which include the two-point and three-point models. The one-point model samples the spectral features of vowels at one point of vowel duration, while the two-point and three-point models sample the features at two and three points of vowel duration. Various combinations of sampled points and predictors (F0, F1, F2, and F3) were analyzed, and the combinations with the best classification accuracy were compared across the models. The results showed that the steady-state model showed the highest classification accuracy when the spectral features and fundamental frequency were sampled at 50% of vowel duration, while the trajectory models showed the highest classification when sampled at 30% and 70% and 10%, 50%, and 90% for two-point and three-point models, respectively. Classification performance was the highest for all models when all parameters (F0, F1, F2, F3) were included across all models. When compared across the models, the trajectory models perform better than the steady-state model. In addition, vowel duration as a parameter has facilitated the classification accuracy for specific vowels. This paper obtains additional evidence for VISC in vowel classification, including detailed classification results of each vowel, identifying the misclassified vowels, and providing insights for vowel classification models.

KEYWORDS

vowel classification, vowel dynamics, discriminant analysis, conversational speech corpus

1. Introduction

Increasing attention has been given to dynamic cues, specifically vowel inherent spectral change (VISC), in addition to the first (F1) and second (F2) formants, which are crucial in classifying the vowels (e.g., Hillenbrand et al. 1995, Morrison 2013, Nearey and Assmann 1986, Peterson and Barney 1952, Zahorian and Jagharghi 1993). The trajectory patterns included these formants measured at both the onset and offset of the vowel in an experimental framework designed to elicit citation forms. These trajectory models evaluated the spectral features from two or three distinct points: onset and offset, or onset, steady state, and offset. These trajectory models are compared to the steady-state model, which only includes the midpoint of a vowel, demonstrating that the trajectory models are superior at classifying vowels. This study implements a supervised learning technique, quadratic discriminant analysis, to examine whether the exceptional performance of the two-point and three-point trajectory models can be generalized to a conversational speech corpus.

Hillenbrand et al. (1995), one of the exemplary studies, demonstrated that the two-point model outperformed the steady-state model, the one-point model, in vowel classification using quadratic discriminant analysis. This study sampled the spectral values in an /hVd/ syllable obtained from English speakers. The two-point model sampled the formant values at 20% and 80% of vowel duration, whereas the one-point model sampled the values at 50% of vowel duration. The classification accuracy indicated that the vowel dynamics contributed to classification by substantially enhancing the combination of F1 and F2. When all of the spectral features, F1, F2, and F3, along with the fundamental frequency (F0), were employed as predictors, the accuracy rate of the classification also escalated. The three-point model, which includes the spectral values sampled at 20%, 50%, and 80% of vowel duration, yielded classification results comparable to those of the two-point model. Hillenbrand et al. (2001) showed analogous results, in which the formant values of the two-point model were sampled at 20% and 70% of vowel duration. In this study, the formant values were collected from CVC syllables, specifically the 'stop+vowel+stop' syllable from native English speakers. The classification accuracy utilizing a quadratic discriminant classifier showed higher accuracy rates in the two-point model than in the one-point model. Both studies demonstrated escalated classification accuracy in the trajectory models regardless of the inclusion of vowel duration as a predictor. These studies illustrated that vowel dynamics provide more crucial information in vowel identification than the steady state.

Research on other languages also corroborates the superior performance of trajectory models in vowel classification. Adank et al. (2004) examined 15 Northern and Southern Standard Dutch vowels using the /sVs/ syllable. Quadratic discriminant analysis indicated that the two-point model, which also incorporated vowel duration as a parameter, performed better, wherein the spectral values were sampled at 25% and 75% of vowel duration. A recent study on Hijazi Arabic demonstrated similar results (Almurashi et al. 2024). This study collected data from monosyllabic or disyllabic words produced in phrases. Discriminant analysis showed that the two-point model identified the vowels with higher accuracy than the one-point model in various F1, F2, F3, and F0 combinations. The study also analyzed the three-point and seven-point models, where the formant values were sampled at 20%, 50%, and 80% in the former and every 10% interval from 20% to 80% of vowel duration in the latter. While the three-point model showed similar results to the two-point model, the seven-point model performed significantly better.

In contrast to these studies that employed the citation forms, Hong (2021, 2023) focused on a conversational speech corpus, the Seoul Corpus, of native Korean speakers, utilizing a neural network model. In Hong (2021), the four-point model, which sampled formant values at 20%, 40%, 60%, and 80% of vowel duration, achieved the best performance. Hong (2023) focuses on the offset, sampling after the midpoint. The one-point model in this

study sampled formant values at 100% of vowel duration. The study did not include the onset values before the midpoint in the trajectory models. Instead, the trajectory models incorporated combinations of spectral values sampled from 100% to 50%. Despite not including the onset, the study demonstrated that the trajectory models still performed better than the one-point model. These studies show that the trajectory models outperform the steady-state model in conversational speech as well.

Interestingly, the trajectory models do not consistently outperform the steady-state model. Vowel dynamics have different influences on the classification of monophthongs and diphthongs. Harrington and Cassidy (1994) studied Australian English vowels in the context of /CVd/ syllables in citation form. The one-point and three-point models, which sampled the formant values at 20%, 50%, and 80% of vowel duration, were analyzed using Gaussian and neural network classification techniques. The results showed that while dynamic cues are critical for vowel identification, they play a more substantial role in diphthongs than monophthongs. In other words, the steady state remains a significant predictor for classifying monophthongs in citation form. In a study by Neel (2004), which involved human subjects performing a vowel identification task, the two-point model performed worse than the one-point model. The two-point model sampled formant values at 10% and 90% of vowel duration. The study also compared the three-point, five-point, and eleven-point models, which sampled formant values at 10%, 50%, and 90%, at 20% intervals from 10% to 90%, and at 10% intervals from 0% to 100% of vowel duration, respectively. While the two-point model showed lower accuracy than the one-point model, the classification accuracy improved with other trajectory models, the three-point, five-point, and eleven-point models. These findings suggest that although the trajectory models are generally better suited for vowel classification, the specific sampling points of formant values are also crucial for performance.

While the dynamic cues of vowels are essential in vowel classification, it remains uncertain whether the results can be generalized to natural conversational speech. Previous studies have employed citation forms such as /hVd/ or /CVC/ within controlled experimental settings, resulting in clear outcomes. However, as Hillenbrand (2013) points out, “the situation may not be this simple with connected speech and more complex phonetic environments” (p. 16). Although real-world speech includes a significantly greater degree of variability that could potentially influence the findings, limited research has been conducted on dynamic cues in natural speech. Therefore, this paper aims to expand the findings, highlighting the importance of vowel dynamics to natural conversational speech. This research seeks to connect experimental data and real-world speech by utilizing the Buckeye Corpus of conversational speech (Pitt et al. 2007). To explore this, it compares the classification performance of the one-point, two-point, and three-point models. The two-point and three-point models were selected for the trajectory models, as there is limited research on the topic concerning spontaneous speech. The traditional onset and offset model was employed for the two-point model, while the onset, steady state, and offset model was utilized for the three-point model. Specifically, this study examines if the trajectory models, the two-point and three-point models, exhibit superior classification performance compared to the one-point model within the domain of conversational speech data. Each model considers varying sampling points of formant values, reflecting the various combinations of sampling points used in previous studies. To address this question, this paper employs a supervised machine learning technique of discriminant analysis, specifically quadratic discriminant analysis (hereafter, QDA). The classification accuracy results indicate that the trajectory models outperform the one-point model when all formants and F0 are involved as predictors, extending previous findings to natural conversational speech. In addition, classification accuracy increased when vowel duration was added as a predictor. By addressing whether vowel dynamics in conversational speech can improve classification accuracy, this study enhances our understanding of vowel acoustics. It has potential implications for improving automatic speech recognition systems that rely on natural speech.

The paper is structured as follows: Section 2 presents methodologies concerning the data and analysis. Section 3 presents the results, covering both the acoustic analyses and the outcomes of QDA. Classification accuracy results are provided for each model and compared across the models. In addition, results with vowel duration as a predictor will be presented. Section 4 discusses the implications of the results and provides concluding remarks.

2. Methods

2.1 Data

The Buckeye Corpus of conversational speech comprises spontaneous speech from 40 adults (20 female) in central Ohio, USA (Pitt et al. 2005, 2007). The corpus comprises two age groups, one under 30 and the other over 40, collected from middle-class Caucasians. The participants took part in the interview, which lasted from 30 to 60 minutes. The corpus includes about 300,000 tokens with phonemic labels, showing an overall 80% agreement among transcribers. Vowels include monophthongs, diphthongs, nasalized vowels, and syllabic consonants. For this study, this paper includes only eight monophthongs. Since the corpus coded [ə] and [ʌ] as one symbol, 'ah', the vowels were excluded from the data. Thus, eight monophthongs were used for analysis. The corresponding symbols used in the corpus data and IPA symbols are shown in Table 1.

Table 1. Symbols Used in the Buckeye Corpus and Corresponding IPA Symbol

Buckeye symbol:	iy	ih	eh	ae	uw	uh	ao	aa
IPA:	i	ɪ	ɛ	æ	u	ʊ	ɔ	ɑ

Nonverbal labels such as 'laughter' were excluded, as were the values that failed to be extracted by the Praat script (Boersma and Weenink 2023, for the script, Yoon 2021). The script extracted a total of 57,501 vowel tokens. The specific number of each vowel and their proportions are shown in Table 2.

Table 2. The Number and Proportions, which are Rounded to One Decimal Point, of Vowels Tokens

iy	ih	eh	ae	uw	uh	ao	aa	Total
7051	18853	12763	5419	2499	2451	3546	5225	57501
12.3%	32.8%	22.2%	9.0%	4.3%	4.2%	6.2%	9.1%	100%

2.2 Acoustic and Discriminant Analyses

For the acoustic analysis, the fundamental frequency (F0) and the three lowest formant frequencies (F1, F2, F3) were extracted by Praat (Boersma and Weenink 2023) using modified scripts from Yoon (2021). Formant and fundamental frequencies were sampled at 10% intervals across the vowel duration. The formant extraction parameters were set to maximum formant values of 5,000Hz for male speakers and 5,500Hz for female speakers, with a window size of 1s. Linear formant frequencies in Hz were sufficient for vowel classification as vowel normalization techniques did not significantly improve the classification accuracy (Hillenbrand and Gayvert 1993, Hillenbrand et al. 1995, Hong 2021). Therefore, linear formant frequencies were utilized for the analyses.

Steady-state and trajectory models were based on formants and F0 combinations. In all models, the analyses were conducted on four sets of predictors: (i) F1 and F2, (ii) F1, F2, and F3, (iii) F1, F2, and F0, and (iv) F1, F2, F3, and F0. Each predictor combination was applied to the one-point, two-point, and three-point models. In the

one-point model, the predictors included the values sampled at a single point from 10% to 90% of vowel duration at 10% intervals. In the two-point model, the values were sampled at two different time points of vowel duration, yielding four different sets: (i) 10% and 90%, (ii) 20% and 80%, (iii) 30% and 70%, and (iv) 40% and 60%. Similarly, the three-point model included the values from three different time points of vowel duration, incorporating the midpoint (50%) along with the predictors in the two-point model. Thus, the combinations were (i) 10%, 50%, and 90%, (ii) 20%, 50%, and 80%, (iii) 30%, 50%, and 70%, and (iv) 40%, 50%, and 60%. These combinations in the two-point and three-point models are represented with a '+' sign, summarized in (1).

(1) a. One-point model:
F0 and the three lowest formants were sampled at a single time point during vowel duration: at 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90% of vowel duration.

b. Two-point model:
F0 and the three lowest formants were sampled at two different time points during vowel duration: at 10%+90%; 20%+80%; 30%+70%; and 40%+60% of vowel duration.

c. Three-point model:
F0 and the three lowest formants were sampled at three different time points during vowel duration: at 10%+50%+90%; 20%+50%+80%; 30%+50%+70%; and 40%+50%+60% of vowel duration.

For discriminant analysis, the data set was split into a training set comprising 70% of the total data and a test set comprising 30% of the data. Each model was trained using the training set and then evaluated to determine whether it correctly predicted the vowel classes using the test set. Table 3 shows the number of tokens and the proportions of each vowel in the training and test sets.

Table 3. The Number and the Proportions of Vowel Tokens Used in the QDA

	iy	ih	eh	ae	uw	uh	ao	aa	Total
Train	4936	13198	8935	3605	1750	1691	2483	3658	40256
Test	2115	5655	3828	1544	749	724	1063	1567	17245
Proportions	12.3%	32.8%	22.2%	9.0%	4.3%	4.2%	6.2%	9.1%	100%

Note: The proportions are rounded up to 1 decimal point.

The table shows that the vowel classes are unbalanced, with 'ih' and 'eh' having the highest proportions compared to the vowels 'uw' and 'uh'. To avoid bias based on the frequency during the training, the training items with fewer tokens were upsampled to match the highest number of token counts. In other words, the number of vowel tokens for the other vowels was increased to align the token count of the vowel 'ih'. Therefore, a total of 105,584 tokens were used in the training (13,198 tokens × 8 vowels). Each model was trained with the balanced training set and then tested on the unbalanced test set, which reflects the distribution of vowel classes in the real world to predict the accuracy of vowel classification. Discriminant analysis implemented QDA, a supervised machine learning algorithm, in R (R Core Team 2023) using the 'caret' package (Kuhn 2008). The QDA was performed with 10-fold cross-validation to evaluate the model performance.

3. Results

3.1 Vowel Trajectory and Formant Analysis

As the lowest three formants were taken every 10% interval of vowel duration, trajectory patterns of the formants are visualized in Figures 1 and 2. Figures show each subject's average values of formants (the lighter lines) and the overall average of each formant for each vowel across subjects (the thick solid lines). The trajectory patterns for both males and females are similar in front and back vowels, while the frequency range is slightly lower for males, as expected.

In addition to the vowel trajectory, the trajectory pattern of F1 and F2 is plotted on the F1-F2 space, as shown in Figure 3. F1 and F2 are averaged across the subjects at every 10% interval of vowel duration. Each circled dot represents 10% intervals of the vowel duration from 10% to 90%, where the triangle represents the offset. As expected, the formant frequencies of females are higher in larger spaces. The trajectory pattern of the vowels across females and males shows similarity in movement patterns. All vowels show a scoop-like movement, indicated by the up-down movement along F1, except the two vowels, 'iy' and 'uw', which move front-back along F2.

The distribution of the vowels at the midpoint of vowel duration is shown in Figure 4. Each label represents an average of each subject, and the solid dots represent an average of each vowel across subjects. While the overall distribution of males and females shows similarities, one noticeable distribution of the vowels is that the back non-high vowels 'ao' and 'aa' are distinctly apart from the other vowels. Also, it is noticeable that the high vowels 'iy', 'ih', 'uw', and 'uh' overlap (except 'iy' of females). The distribution of females is broader and higher than males, as expected.

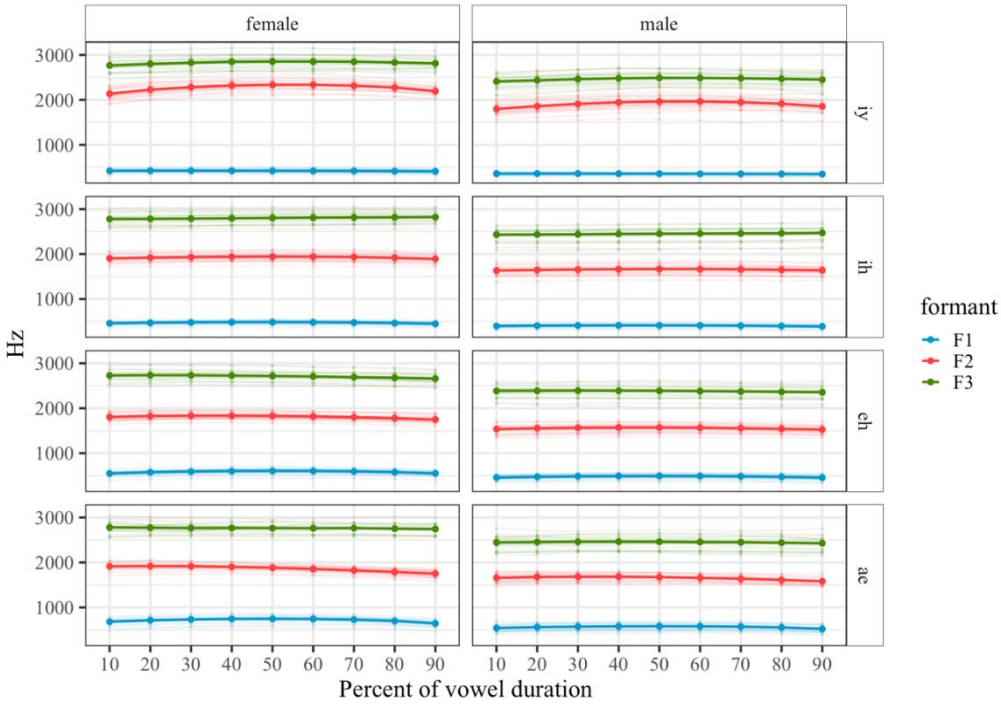

Figure 1. Vowel Trajectory of Front Vowels. (*Note*: A thick solid line represents the average of the vowels, and thin lines represent the average of each subject.)

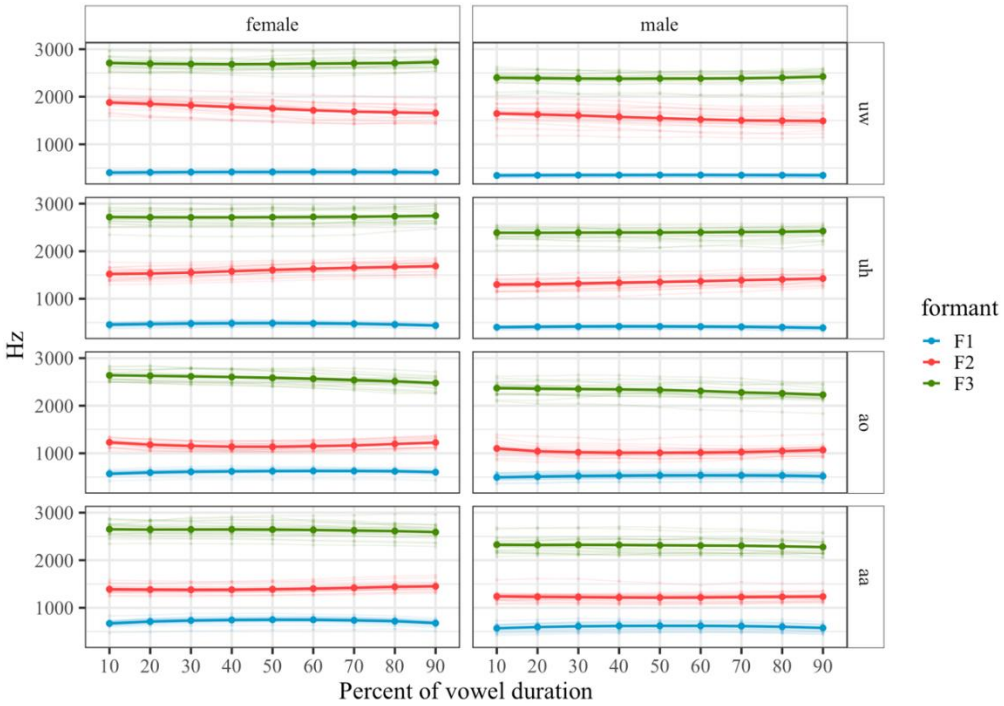

Figure 2. Vowel Trajectory of Back Vowels. (*Note*: The thick solid line represents the average of each vowel, and the thin lines represent the average of each subject in each vowel.)

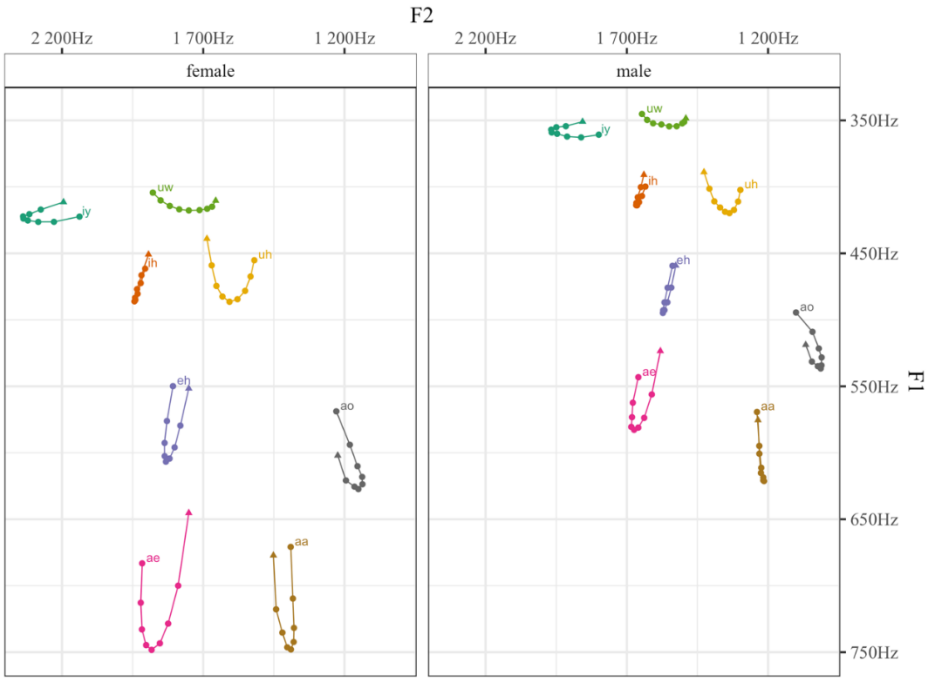

Figure 3. Vowel Trajectory of Females and Males on F1-F2 Space. (*Note*: Each dot represents 10% intervals of vowel duration from 10% to 90%. The triangle indicates the offset (90%).)

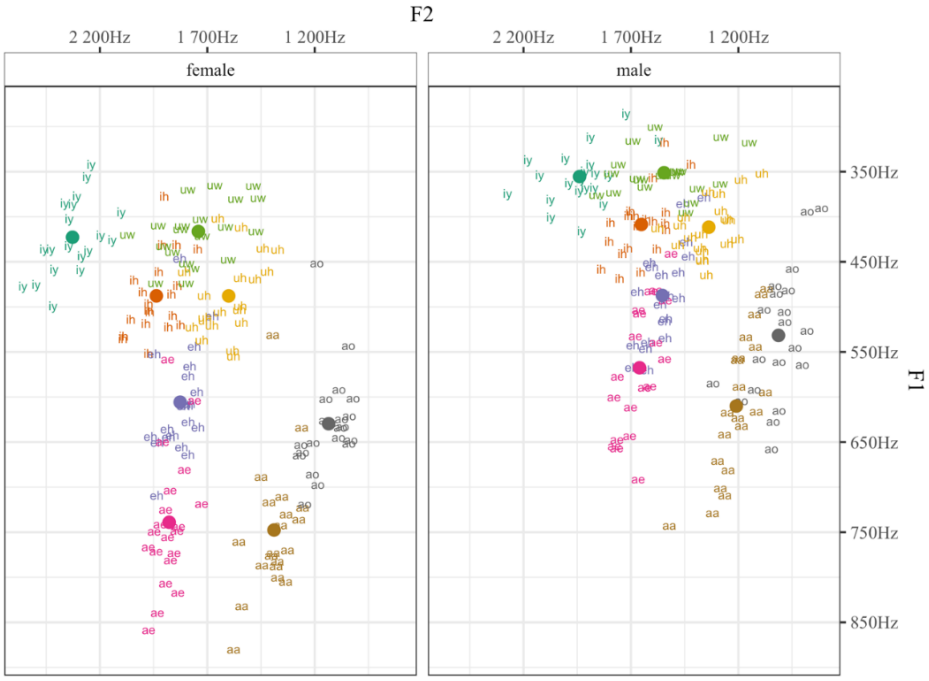

Figure 4. Distribution of Vowels Measured at the Midpoint of the Vowel Duration. (*Note*: Each label represents each subject's average, and the dots represent the average across the subjects.)

The mean of vowel duration of each vowel by sex and overall mean is shown in Table 4. The low vowels 'ae', 'ao', and 'aa' show relatively longer duration, and the tense vowels 'iy' and 'uw' are longer than the counterparts 'ih' and 'uh', respectively. The mid-front lax vowel 'eh' is shorter than 'iy' and 'uw' but longer than 'ih' and 'uh'. Interestingly, 'iy' and 'uw' show similar durations, and 'ih' and 'uh' show similar durations.

Table 4. Mean Values of Vowel Duration of Each Vowel

	iy	ih	eh	ae	uw	uh	ao	aa
Female	90.97	62.21	81.86	117.42	90.11	62.98	106.02	106.94
Male	89.73	63.89	76.17	119.75	89.23	63.44	110.06	103.33
Mean	90.35	63.05	79.02	118.59	89.67	63.21	108.04	105.14

3.2 Discriminant Analysis

Quadratic discriminant analysis (QDA) was conducted on each model, the one-point, two-point, and three-point models, with various combinations of formant frequencies and fundamental frequency sampled at different time slices as predictor variables. QDA was conducted on the training set, and then the model predicted the classification accuracy on the test set. The overall accuracy rate for vowel classification of each model was calculated from true positives, that is, the proportion of the actual vowels that are correctly predicted by the model based on the total occurrences of vowels. In addition to the overall accuracy, the confusion matrix with the highest accuracy rate of the test set is presented to understand the correct classification and misclassifications in detail. Based on the confusion matrix, recall and precision are calculated. Recall measures the proportion of how well the model

identifies the true positives, indicating that high recall is effective at classifying the actual vowel.[1] Precision measures the proportion of positive identifications that are actually correct, in other words, how reliable the predictions are for each class.[2] High precision, therefore, indicates that the model predicts a vowel and is often correct.

As the distribution of vowel classes in the test set is imbalanced, the F1 scores are evaluated. The F1 score represents the harmonic mean of recall and precision, offering a balanced measure that considers both (Kelleher et al. 2020).[3] That is, the F1 score considers the true positives, the actual vowels that are correctly predicted by the model, misclassified actual vowels (false negatives), and predicted vowels that are incorrectly classified (false positives). A high F1 score is achieved only when both recall and precision are high, ensuring a balanced performance in correctly identifying vowels while avoiding false positives and negatives. This provides a more robust evaluation of the model's ability to correctly classify all vowel categories, regardless of their frequency in the dataset, which is crucial in real-world applications where class imbalance is common.

3.2.1 One-point model

The one-point model includes combinations of F1, F2, F3, and F0 taken at one point within vowel duration as predictors. The input of the model was the combinations sampled at every 10% interval from 10% to 90% of vowel duration. The overall accuracy of classification for the training set is provided in Table 5, where the numbers are rounded up to three decimal points.

Table 5. Classification Accuracy of the One-Point Model with Training Set

	10%	20%	30%	40%	50%	60%	70%	80%	90%
F1, F2	0.362	0.405	0.431	0.452	0.461	0.457	0.443	0.417	0.380
F1, F2, F3	0.382	0.422	0.455	0.476	0.484	0.482	0.468	0.441	0.402
F0, F1, F2	0.379	0.429	0.461	0.482	0.490	0.485	0.471	0.440	0.397
F0, F1, F2, F3	0.390	0.436	0.470	0.495	0.504	0.502	0.485	0.454	0.415

In all the combinations of formants and F0, classification accuracy increases towards the midpoint while the accuracy decreases towards the edges of vowel duration. The results support that the midpoint, 50% of the vowel duration, is crucial in vowel identification. Also, including either F3 or F0, or both, increase the accuracy compared to F1 and F2 combination as a predictor. With the training set, the midpoint of vowels as a predictor shows the highest classification accuracy with the F1, F2, F3, and F0 combination (0.504).

[1] $Recall = \frac{True\ Positive}{True\ Positive+False\ Negative}$. As mentioned in the text, true positives refer to the cases where the model correctly predicts the actual vowels, while false negatives refer to the cases where the model incorrectly identifies the actual vowel as other vowels. For example, if a model predicts a vowel as 'iy' and when it is actually 'iy', this is a true positive case. False negative is when the model fails to identify the vowel as 'iy' when it is actually 'iy'.

[2] $Precision = \frac{True\ Positive}{True\ Positive+False\ Positive}$. False positive refers to the cases where the model incorrectly identifies a vowel that is not. For example, the model incorrectly predicts a vowel as 'iy' when it is actually a different vowel.

[3] $F1 = 2 \times \frac{Precesion \times Recall}{Precesion+Recall}$

The same model predicted the classification with the test set, which keeps the proportions of the original data set. The boldfaced numbers indicate a significant difference from the no information rate.[4] Classification results with the test set are in Table 6.

Table 6. Classification Accuracy of the One-Point Model with Test Set

	10%	20%	30%	40%	50%	60%	70%	80%	90%
F1, F2	0.293	0.337	**0.374**	**0.397**	**0.400**	**0.396**	**0.379**	**0.347**	0.295
F1, F2, F3	0.318	**0.365**	**0.404**	**0.425**	**0.433**	**0.427**	**0.409**	**0.375**	**0.330**
F0, F1, F2	0.312	**0.366**	**0.405**	**0.427**	**0.436**	**0.432**	**0.409**	**0.371**	0.319
F0, F1, F2, F3	0.330	**0.377**	**0.418**	**0.442**	**0.447**	**0.445**	**0.423**	**0.388**	0.338

Note: The boldfaced numbers indicate that the accuracy is significantly different from the no information rate.

Note that none of the predictors sampled at 10% of vowel duration classify the vowels, which indicates that the formant and F0 values sampled at the beginning of the vowel alone are not informative for vowel identification. At 20% of vowel duration, the combination of F1 and F2 does not classify vowels significantly, whereas at 90% of vowel duration, only the combination of F1, F2, and F3 classifies the vowels significantly. The results indicate that when the formant and F0 values are sampled at the edges of vowels, the model does not learn to classify the vowels only with F1 and F2. The highest accuracy result also comes from the midpoint with F1, F2, F3, and F0 as the training set. Therefore, the accuracy rate of 50% in both training and the test will be used as the baseline for this paper and compared to other predictors in the two-point and three-point models.

Table 7 is a confusion matrix of the test set with F1, F2, F3, and F0 sampled at 50% of vowel duration. The boldfaced numbers indicate the actual vowels correctly classified by the model, the true positive cases. As mentioned above, precision refers to how well the model predicts the actual vowel, and recall refers to the actual vowels that are correctly classified as the vowel. In terms of recall, the vowels ‘iy’, ‘aa’, ‘ao’, and ‘ae’ show relatively high rates, which indicates that the model correctly identifies the vowel when it is actually the vowel, whereas the vowels ‘ih’, ‘uh’, ‘uw’, and ‘eh’ show low recall rates. The model especially rarely identifies the vowel ‘ih’, showing a very low recall (0.242). As Table 7 shows, the vowel ‘ih’ is often misclassified as ‘iy’, ‘uw’, and ‘eh’, by 1,232 times, 1,013 times, and 937 times, respectively, which overlap in the vowel space. Likewise, the vowel ‘uh’ is often misclassified as ‘uw’ by 170 times, ‘uw’ as ‘iy’ by 207 times, and ‘eh’ as ‘ae’ by 614 times. The model effectively detects the vowels ‘iy’, ‘aa’, ‘ao’, and ‘ae’, whereas it misses a great proportion of the vowels ‘ih’, ‘uh’, ‘uw’, and ‘eh’. Precision shows some interesting performance. Although other vowels show moderate performance in precision, the vowels ‘uw’ and ‘uh’ show very low precision rates. The model misclassifies the vowel ‘uw’ as ‘ih’ the most, 1013 times, followed by ‘eh’ 227 times. The model also misclassifies the vowel ‘uh’ as ‘ih’ and ‘eh’ by 776 and 315 times, respectively. That is, the model struggles to identify the vowels ‘uw’ and ‘uh’ by showing very low precision and low recall, which means that the vowels are often misidentified as other vowels, especially as ‘ih’.

[4] The no information rate refers to the accuracy that could be achieved by predicting the most frequent class in a dataset. Therefore, if the accuracy result is significantly higher than the no information rate ($p < 0.05$), the model has learned classification patterns from the predictors that are better than simply guessing the most common class.

Table 7. Confusion Matrix of the Test Set with F0, F1, F2, and F3 Sampled at 50% of Vowel Duration as a Predictor in the One-Point Model

		Actual								Precision
		iy	ih	eh	ae	uw	uh	ao	aa	
Prediction	iy	**1618**	1232	171	111	207	40	5	2	0.478
	ih	120	**1369**	490	81	48	69	3	1	0.628
	eh	86	937	**1614**	329	31	75	26	74	0.509
	ae	14	140	614	**840**	5	7	30	135	0.471
	uw	208	1013	227	26	**299**	170	23	21	0.150
	uh	40	776	315	34	95	**267**	51	73	0.162
	ao	21	96	99	9	61	78	**682**	241	0.530
	aa	8	92	298	114	3	18	243	**1020**	0.568
Recall		0.765	0.242	0.422	0.544	0.399	0.369	0.642	0.651	

Note: The boldfaced numbers indicate the true positive cases.

Overall, the predictors, F0, F1, F2, and F3 sampled at the midpoint, show the highest accuracy result in the one-point model. The one-point model classifies well with the vowels ‘iy’, ‘ao’, and ‘aa’ showing decent recall. However, the model struggles with the vowels ‘ih’, ‘uw’, and ‘uh’, showing low recall and precision, which indicates that these vowels are often misidentified as other vowels.

3.2.2 Two-point model

The two-point model includes combinations of formant and fundamental frequencies sampled at two points of the vowel duration as predictors. The two-time slices were paired as one point before the midpoint and the other after the midpoint. The pairs are 10%+90%, 20%+80%, 30%+70%, and 40%+60%. The F1, F2, F3, and F0 combinations sampled at these two points were the predictors in the two-point model. The overall classification accuracy with the training set is shown in Table 8. The results show that the classification accuracy in all combinations of the predictors is the highest at 30%+70% and the lowest at 10%+90%. Specifically, 20%+80% and 40%+60% show similar results in the combinations of F1, F2 and F0, F1, F2, in that 0.001 higher in 40%+60% of F1, F2 and 0.002 higher in 20%+80% in F0, F1, F2. When F3 is included, 20%+80% show higher classification result, 0.011 in F1, F2, F3 and 0.013 in F0, F1, F2, F3.

Table 8. Classification Accuracy of the Two-Point Model with the Training Set

	10%+90%	20%+80%	30%+70%	40%+60%
F1, F2	0.441	0.473	0.487	0.474
F1, F2, F3	0.476	0.504	0.515	0.493
F0, F1, F2	0.468	0.500	0.511	0.498
F0, F1, F2, F3	0.492	0.520	0.530	0.507

Compared to the classification accuracy of the one-point model (50%) with the training set (Table 5), all combinations of the two-point models showed higher classification accuracy except 10%+90%. The higher accuracy rate of classification of the two-point model aligns with other trajectory studies (Hillenbrand et al. 1995, for example), which show that the trajectory information facilitates the higher classification rate. The lower classification accuracy of 10%+90% is consistent with Neel (2004), who showed that the one-point model performs better than the two-point model since formants of the edges of vowels do not provide the “target” frequencies to identify the vowels.

The same model predicted classification with the test set; the results are shown in Table 9. The boldfaced fonts refer to a significant difference from the no information rate. In the two-point model, only the combination of F1 and F2 sampled at 10% and 90% was not statistically significantly different from the no information rate. The overall classification accuracy of the two-point model demonstrates that the 30%+70% condition shows the highest and 10%+90% the lowest classification accuracy as the training set. Unlike the training set, the classification accuracy in 40%+60% is consistently higher than in 20%+80% across all combinations of predictors: 0.014 higher in F1, F2, 0.010 higher in F1, F2, F3, 0.020 higher in F0, F1, F2, and 0.012 higher in F0, F1, F2, F3.

Table 9. Classification Accuracy of the Two-Point Model with Test Set

	10%+90%	20%+80%	30%+70%	40%+60%
F1, F2	0.335	**0.368**	**0.399**	**0.382**
F1, F2, F3	**0.376**	**0.405**	**0.428**	**0.415**
F0, F1, F2	**0.365**	**0.403**	**0.446**	**0.423**
F0, F1, F2, F3	**0.389**	**0.428**	**0.459**	**0.440**

Note: The boldfaced numbers indicate that the accuracy is significantly different from the no information rate.

The confusion matrix with F1, F2, F3, and F0 sampled at 30%+70% is shown in Table 10, which demonstrates the increase of the classification accuracy resulting from the higher correct classification of the vowels ‘uw’ and ‘uh’, that is, higher recall and precision in the vowels. Regarding recall, the vowels ‘iy’, ‘aa’, ‘ao’, and ‘ae’ show relatively high rates as in the one-point model. Like the one-point model, the vowel ‘ih’ shows the lowest recall (0.248) and misclassifies the vowel with ‘iy’ by 1,253 times. The vowel is also misclassified as ‘uw’, ‘eh’, and ‘uh’ by 974 times, 864 times, and 830 times, respectively, indicating that the model struggles with the vowel ‘ih’. However, unlike the one-point model, the two-point model shows a notable increase in the vowels ‘uw’ and ‘uh’, which increased by 0.1 and 0.151, respectively. In other words, the vowels ‘uw’ and ‘uh’ are more correctly identified in the two-point model. Precision slightly increased in all vowels, except the vowel ‘ih’, which decreased by 0.012. The performance of vowels ‘uw’ and ‘uh’ notably increased by 0.41 and 0.412, respectively, although precision is low. The model still misclassifies ‘uw’ and ‘uh’ as ‘ih’ the most, 974 times and 830 times, respectively. In other words, although the two-point model struggles to identify the vowels ‘uw’ and ‘uh’ by showing low precision, the model is more accurate in predicting the vowels than the one-point model. These results indicate that the increase in the overall accuracy rate is due to better classification of the two high back vowels, ‘uw’ and ‘uh’.

Table 10. Confusion Matrix of the Test Set in the Two-Point Model with F1, F2, F3, and F0 Sampled at 30% and 70% of the Vowel Duration as Predictors

	Actual								Precision
Prediction	iy	ih	eh	ae	uw	uh	ao	aa	
iy	**1589**	1253	173	66	179	39	4	1	0.481
ih	149	**1404**	496	90	47	78	7	8	0.616
eh	88	864	**1716**	420	27	53	27	100	0.521
ae	21	149	495	**775**	13	9	18	136	0.480
uw	197	974	204	28	**374**	115	44	17	0.192
uh	53	830	352	35	80	**376**	52	75	0.203
ao	13	96	140	26	23	39	**659**	206	0.548
aa	5	85	252	104	6	15	252	**1024**	0.587
Recall	0.751	0.248	0.448	0.502	0.499	0.519	0.620	0.653	

Note: The boldfaced numbers are the true positive cases.

Overall, the two-point model, the trajectory model classifies the vowels ‘uw’ and ‘uh’ more accurately than the one-point model by performing better in recall and precision, while other vowels show similar results. The results in the two-point model show that the trajectory pattern is a better indicator for certain vowels. The model still struggles with the vowel ‘ih’, which shows low recall, which indicates that it is often misidentified as other vowels.

3.2.3 Three-point model

The three-point model includes combinations of formant frequencies and F0 sampled at three points as predictors. The midpoint was added to the predictors in the two-point model, producing four predictors: 10%+50%+90%, 20%+50%+80%, 30%+50%+70%, and 40%50%+60%. The overall classification accuracy with the training set is shown in Table 11. Classification accuracy in all combinations of formants and F0 is the highest at 10%+50%+90%, followed by 20%+50%+80%, 30%+50%+70%, and 40%+50%+60%. Contrary to the two-point model, predictors in 10%+50%+90% showed the highest accuracy.

Table 11. Classification Accuracy of the Three-Point Model with the Training Set

	10%+50%+90%	20%+50%+80%	30%+50%+70%	40%+50%+60%
F1, F2	0.499	0.484	0.465	0.416
F1, F2, F3	0.527	0.506	0.479	0.418
F0, F1, F2	0.518	0.504	0.480	0.414
F0, F1, F2, F3	0.537	0.518	0.489	0.419

Compared to the training set results of the one-point model (Table 5), all predictors in 10%+50%+90% and 20%+50%+80% show higher classification accuracy. Especially, classification accuracy rate increased largely in 10%+50%+90%; 0.038 in F1, F2, 0.039 in F1, F2, F3, 0.028 in F0, F1, F2, and 0.032 in F0, F1, F2, F3. In 30%+50%+70%, only the combination of F1 and F2 showed a slight increase, by 0.004, while other predictors decreased in accuracy and the predictors in 40%+50%+60%. Adding temporal information to the midpoint increases the classification accuracy, while the temporal information near the midpoint does not facilitate the classification of the vowels.

Compared to the counterparts in the two-point model, adding the midpoint increased classification accuracy notably in 10%+50%+90%, while 40%+50%+60% show a decrease. In 10%+50%+90%, accuracy has increased 0.057 in F1, F2, 0.048 in F1, F2, F3, 0.05 in F0, F1, F2, and 0.044 in F0, F1, F2, F3. The counterpart in the two-point model, 10%+90%, performed the lowest among the predictors in the two-point model; however, adding the midpoint to the trajectory pattern has significantly increased accuracy across all combinations of the formants and F0. On the contrary, accuracy decreased across all predictors in 40%+50%+60% compared to the counterpart in the two-point model. The results are interesting in that adding the midpoint toward the edges of the vowel duration notably increases classification accuracy, while adding the same information closer to the midpoint results in the opposite direction. The findings indicate that the vowel trajectory patterns are better captured in 10%+50%+90% than in 40%+50%+60%, which do not reflect the full trajectory patterns of the vowels. Therefore, adding the midpoint to nearby points is not informative for vowel classification.

The model also predicted classification accuracy on the test set, as shown in Table 12. In the three-point model, the classification accuracy of all the predictors was significantly different from the information rate. The predictors F0, F1, F2, F3 sampled at 10%+50%+90% performed the highest accuracy (0.459) among the predictors. In the three-point model, however, it is not the case that predictors sampled at 10%+50%+90% resulted in the highest

accuracy across all combinations of spectral features. In other words, the accuracy rate shows some mixed results based on the combination of formants and F0 and at which point the spectral features are sampled. The F1, F2 predictor set showed the highest classification accuracy at 10%+50%+90% followed by 40%+50%+60% with a 0.005 difference. When F3 is included as a predictor, 10%+50%+90% showed the highest classification accuracy, followed by 20%+50%+80% with a 0.02 difference. When F0 is included, 40%+50%+60% showed the highest classification accuracy, followed by 30%+50%+70% with a 0.001 difference. When F0 and F3 were included as predictors, 10%+50%+90% showed the highest classification accuracy, followed by 30%+50%+70% with a 0.008 difference. In the three-point model, each trajectory pattern of time points with various combinations of predictors demonstrates varying degrees of accuracy based on the combinations of formants and F0 rather than a specific time point showing consistently high accuracy results.

Table 12. Classification Accuracy of the Three-Point Model with Test Set

	10%+50%+90%	20%+50%+80%	30%+50%+70%	40%+50%+60%
F1, F2	**0.392**	**0.374**	**0.372**	**0.387**
F1, F2, F3	**0.416**	**0.396**	**0.387**	**0.390**
F0, F1, F2	**0.449**	**0.432**	**0.458**	**0.459**
F0, F1, F2, F3	**0.459**	**0.438**	**0.451**	**0.443**

Note: All test results are significantly different from the no information rate.

Table 13 shows the confusion matrix of the test set with F0, F1, F2, and F3 sampled at 10%+50%+90%. It shows the better classification of the vowels ‘uw’ and ‘eh’, followed by ‘uh’, by showing higher recall than the one-point and two-point models. The actual vowel ‘uw’ is correctly classified as the vowel by 0.645, which increased by 0.246 and 0.146 compared to the one-point and two-point models, respectively. Recall of the vowel ‘eh' has increased by 0.09 and 0.063 compared to the one-point and two-point models, respectively. Additionally, the recall of the vowel ‘uh’ has risen by 0.173 and 0.022 when contrasted with the one-point and two-point models, respectively. Other vowels, except these three, show a slight decrease in recall. The vowels ‘ao’, ‘iy’, and ‘ae’ decreased by 0.091, 0.055, and 0.054 compared to the two-point model and by 0.113, 0.069, and 0.097 compared to the one-point model, respectively. The vowel ‘ih’ showed similar recall to the one-point and two-point models, which showed a 0.001 increase compared to the one-point model and a 0.005 decrease compared to the two-point model. The vowel ‘ih’ is often misclassified as neighboring vowels, such as ‘uw’ by 1,259 times, ‘eh’ by 966 times, and ‘iy’ by 906 times.

Regarding precision, other vowels show moderate performance except for the vowels ‘uw’ and ‘uh’. The vowels ‘uw’ and ‘uh’ are misclassified as ‘ih’ the most by 1,259 times and 851 times, respectively. Although the two vowels show the lowest performance in precision, ‘uw’ and ‘uh’ increased by 0.041 and 0.043, respectively, compared to the one-point model while showing no difference from the two-point model. Compared to the one-point model, precision of other vowels has also increased by 0.066 for ‘iy’, 0.3 for ‘aa’, and 0.28 for ‘ao’. Compared to the two-point model, precision increased in ‘iy’ by 0.063 and ‘ih’ by 0.025. In other words, the model struggles to predict the vowels ‘uw’ and ‘uh’ that are actually correct while predicting other vowels correctly.

Table 13. Confusion Matrix of the Test Set in the Three-Point Model with the Predictor F0, F1, F2, and F3 Taken at 10%, 50%, and 90% of the Vowel Duration

		Actual								Precision
		iy	ih	eh	ae	uw	uh	ao	aa	
Prediction	iy	**1472**	906	142	73	76	28	5	3	0.544
	ih	149	**1377**	410	83	55	60	6	8	0.641
	eh	114	966	**1959**	511	32	56	38	159	0.511
	ae	42	162	403	**691**	10	11	32	122	0.469
	uw	277	1259	253	39	**483**	136	50	23	0.192
	uh	44	851	345	32	67	**392**	94	94	0.204
	ao	14	67	114	26	19	27	**562**	179	0.558
	aa	3	67	202	89	7	14	276	**979**	0.598
Recall		0.696	0.244	0.512	0.448	0.645	0.541	0.529	0.625	

Note: The boldfaced numbers indicate the true positive cases.

Overall, the three-point model classifies the vowels ‘uw’ and ‘uh’ more accurately than the one-point and the two-point models. The results from the three-point model suggest that the trajectory pattern, which includes the midpoint, serves as a better indicator for the classification of specific vowels, ‘uw’ and ‘uh’, as evidenced by a higher recall rate in comparison to other models. However, the model still struggles with the vowel ‘ih’, which is often misclassified with neighboring vowels, resulting in a low recall. Additionally, the three-point model predicts the vowels better than the other two models by showing higher precision; however, it still has difficulties predicting the vowels ‘uw’ and ‘uh’.

3.2.4 Comparison across the models

This section compares the models with the highest accuracy rates with the test set. In addition, the F1 score of each model is compared. As mentioned above, the F1 score is a harmonic mean of recall and precision, a widely used method for comparing the imbalanced distribution of data.

Figure 5 summarizes the test set accuracy results of each model with four sets of predictors that resulted in the highest accuracy. The accuracy for the one-point model is sampled at 50%, for the two-point model at 30% and 70%, and for the three-point model at 10%+50%+90% of the vowel duration. The classification accuracy shows some interesting results compared to the midpoint results in the one-point model (Table 6). With the F1, F2 and F1, F2, F3 predictor sets, the trajectory models show a slightly lower accuracy than the one-point model. While the F1 and F2 predictors in the one-point model show 0.4 and F1, F2, and F3 predictors show 0.433, the accuracy in the three-point model showed 0.392 and 0.416, respectively, which differ by 0.008 and 0.017. However, when F0 was included, the accuracy increased by 0.01 in the two-point model and by 0.013 in the three-point model compared to the one-point model. When both F0 and F3 were included as predictors, classification accuracy increased by 0.012 in the trajectory models compared to the steady-state model. In other words, the lowest three formants alone do not provide enough information for classification in the trajectory models. There could be several reasons for the higher accuracy rate with F0. The analyses utilized the linear formant frequencies, which preserves more information about speaker groups, including sex. F0 is known to differentiate speaker groups (Hillenbrand and Gayvert 1993). Also, each vowel has a different F0; that is, high vowels have a higher F0, while low vowels have a lower F0 (Whalen and Levitt 1995). Therefore, including F0 improved the accuracy rate.

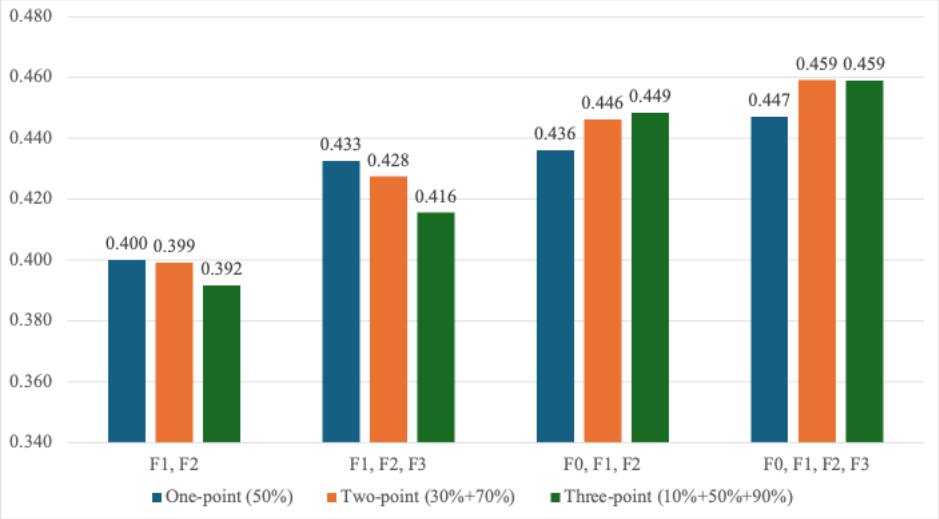

Figure 5. Comparison of Accuracy Rates Across the Models Sampled at Different Time Points for Predictors: 50% in the One-Point Model, 30%+70% in the Two-Point Model, and 10%+50%+90% in the Three-Point Model

Table 14 shows F1 scores for each vowel, which shows different trends depending on the models. The scores are calculated with all predictors, F0, F1, F2, and F3, sampled at each point. In all three models, the vowels 'iy', 'ao', and 'aa' show high scores, indicating that the models correctly identify the actual vowels, and there are a few cases where the vowels are misclassified for one another. The vowels 'uw', 'uh', and 'ih' show low scores across all models. This means that the model is either misidentifying vowels frequently, leading to low precision, failing to detect correct vowels, resulting in low recall, or both. To be more specific, the models frequently predict 'ih' correctly, which leads to high precision, while the actual vowel is misclassified, resulting in low recall. In contrast to 'ih', the models incorrectly identify 'uw' and 'uh' as the vowels, leading to low precision. However, the actual vowels 'uw' and 'uh' are classified correctly, resulting in high recall. Due to the differences in recall and precision, the F1 scores of these vowels are lower than the other vowels. Nonetheless, the vowels 'uw', and 'uh' show a noticeable increase in the trajectory models. The classification accuracy of the vowel 'uw' has increased by 0.058 in the two-point model and 0.077 in the three-point model compared to the one-point model. The vowel 'uh' increased by 0.067 and 0.072 in the trajectory models compared to the steady-state model. The vowels 'ih' and 'eh' also show an increase in the trajectory models. The vowel 'ih' increased by 0.005 in the two-point model, compared to the one-point model, and the vowel 'eh' increased by 0.021 and 0.05 in the two-point and three-point models, respectively. Unlike these vowels, the vowel 'ae' shows some decrease in the trajectory models compared to the one-point model; specifically, the three-point model results in the lowest. The vowel 'ao' shows similar results for the one-point and two-point models but decreases by 0.037 in the three-point model. The results of these two vowels imply that the trajectory pattern is not necessary to classify the vowels for some vowels. The vowel 'aa' shows an increase in the trajectory models, but the score in the three-point model is slightly lower than the two-point model. These results suggest that the models classify the vowels 'iy,' 'ao,' and 'aa' well while struggling with the vowels 'ih,' 'uw,' and 'uh.' However, the trajectory models perform better in the classification of these vowels.

Table 14. The F1 Score in Each Model for Each Vowel

Models	iy	ih	eh	ae	uw	uh	ao	aa
One-point (50%)	0.588	0.349	0.461	0.505	0.219	0.225	0.580	0.607
Two-point (30%+70%)	0.586	0.354	0.482	0.491	0.277	0.292	0.582	0.619
Three-point (10%+50%+90%)	0.611	0.353	0.511	0.458	0.296	0.297	0.543	0.611

Note: Numbers are rounded up to three decimal points.

In summary, the discriminant results show that the model can predict and classify vowels. Among the combinations of the predictors, the F0, F1, F2, and F3 set always showed the highest accuracy in both training and test sets. With the predictor set, the trajectory models, the two-point (sampled at 30%+70% of the vowel duration) and the three-point (sampled at 10%+50%+90% of the vowel duration) models, perform better than the steady-state model, the one-point model (sampled at 50% of the vowel duration). The vowel dynamics are essential in vowel classification, especially for 'ih', 'eh', 'uw' and 'uh'. These vowels are often misclassified as others; however, the trajectory models show a notable increase compared to the one-point model.

3.2.5 Role of vowel duration

In addition to dynamic cues, further analysis was conducted, as vowel duration has been shown to improve vowel classification in many studies (Almurashi et al. 2020, 2024, Hillenbrand et al. 1995, 2001, Hong 2021, Watson and Harrington 1999, Zahorian and Jagharghi 1993). The results of this paper also support the idea that duration is an important parameter for vowel classification, especially for specific vowels.

Table 15 shows the classification accuracy results, which included the vowel duration as an additional predictor for the one-point (50%), two-point (30+70%), and three-point (10%+50%+90%) models for the test set. All the test set results were significantly different from the no information rate, which is indicated by the bold font. With duration, the one-point model resulted in the highest classification accuracy overall, while the three-point model showed the lowest classification accuracy. Numbers in the parentheses indicate the contribution of duration compared to the previous models (Table 6, Table 9, and Table 12), respectively. When duration is included as a predictor in addition to the formants and F0, it has contributed to an increase in accuracy across all models, especially a larger contribution for the one-point model. Within the one-point model, duration facilitates accuracy with F1 and F2 the most. In other words, duration plays a key role in vowel classification when there are no dynamic features for vowels. Therefore, the contribution of duration is the smallest in the three-point model and then the two-point model.

Table 15. Classification Accuracy Results with Vowel Duration Included as a Predictor in Addition to the Formants and Fundamental Frequency.

	One-point (50%)		Two-point (30%+70%)		Three-point (10%+50%+90%)	
F1, F2, dur	**0.471**	(0.071)	**0.452**	(0.053)	**0.426**	(0.035)
F1, F2, F3, dur	**0.498**	(0.065)	**0.473**	(0.045)	**0.449**	(0.033)
F0, F1, F2, dur	**0.498**	(0.062)	**0.496**	(0.050)	**0.489**	(0.040)
F0, F1, F2, F3, dur	**0.508**	(0.061)	**0.507**	(0.048)	**0.502**	(0.035)

Table 16 illustrates the F1 scores of each model when duration is included as a predictor with F0, F1, F2, and F3. In the one-point model, the vowels 'iy', 'aa', and 'ao' show a relatively higher score, and the vowels 'ae', 'ih', and 'eh' show moderate scores, while the vowels 'uw' and 'uh' show lower scores. The two-point and three-point

models show similar tendencies. By adding the dynamic cues, however, the two-point and three-point models showed similar or higher scores in vowels in general compared to the one-point model, except for the vowels 'ih', 'ae', and 'ao'.

Table 16. The F1 Score Comparison Across Models with Duration Added as a Parameter.

Models	iy	ih	eh	ae	uw	uh	ao	aa
One-point (50%)	0.619	0.513	0.481	0.526	0.284	0.272	0.600	0.613
Two-point (30%+70%)	0.618	0.489	0.491	0.524	0.352	0.300	0.589	0.622
Three-point (10%+50%+90%)	0.626	0.465	0.510	0.487	0.344	0.299	0.544	0.619

Note: Numbers are rounded up to three decimal points.

Comparing the models with and without duration shows the role of duration in vowel classification (Table 14 and Table 16, which is visualized in Figure 6). Including duration as a predictor results in similar or higher F1 scores within each model. Mainly, duration affects the classification of the vowel 'ih' across all models and results in a higher F1 score. In the one-point model, the duration increases classification accuracy in vowels 'ih', 'uw', and 'uh' by 0.164, 0.066, and 0.047, respectively. The increase for the vowel 'ih' is remarkable compared to other vowels. In the two-point model, duration facilitates the classification of the vowels 'ih' and 'uw' by 0.135 and 0.075, respectively. Like in the one-point model, the vowel 'ih' shows a remarkable increase with duration. In the three-point model, the duration increases accuracy in the vowel 'ih' by 0.112 and the vowel 'uw' by 0.049. Like the other two models, duration remarkably increases the classification of the vowel 'ih', followed by the vowel 'uw'. In short, the vowel duration is an important factor of classification, mainly for the vowels 'ih' and 'uw'. While duration increases the classification of the two vowels, duration does not affect the vowel 'uh' much. Across all models, whether they include duration or not, 'uh' is frequently misclassified as 'ih', leading to very low precision and, consequently, a low F1 score.[5] One potential explanation for the high rates of misclassification may be the unbalanced dataset, where the vowel 'ih' comprises the largest proportion while 'uh' accounts for one of the smallest proportions. Due to this frequency difference, 'uh' is frequently misclassified as 'ih'. The vowel 'uw' also has low proportions and is often misclassified as 'ih', which results in low F1 scores. Although duration has increased the F1 score for 'uw', these scores are relatively lower compared to those of other vowels. A detailed analysis to adequately distinguish between 'ih', 'uh', and 'uw' requires further investigation.

To sum up, vowel duration facilitates vowel classification across all models, with a more significant impact in the steady-state model compared to the trajectory models. Duration plays a crucial role when there is only F1 and F2 information for the classification. When there is dynamic information, other values can compensate for the role of duration. Thus, the increase is not as significant as the steady-state model. In addition, duration is affected differently by the vowels. Some vowels, such as 'ao' or 'aa', are not greatly affected by duration, while the vowels 'ih' and 'uw' are significantly influenced by vowel duration. It seems that high vowels are more affected or classified better with duration, while low back vowels are less affected by duration. That is, vowel duration facilitates vowel classification overall, and when there are no dynamic cues, duration plays an important role in classification, especially in certain vowels.

[5] Analyses of the confusion matrix with duration were also conducted for all models, and the results show a similar trend to those from the models without duration; therefore, the detailed table is omitted from this paper.

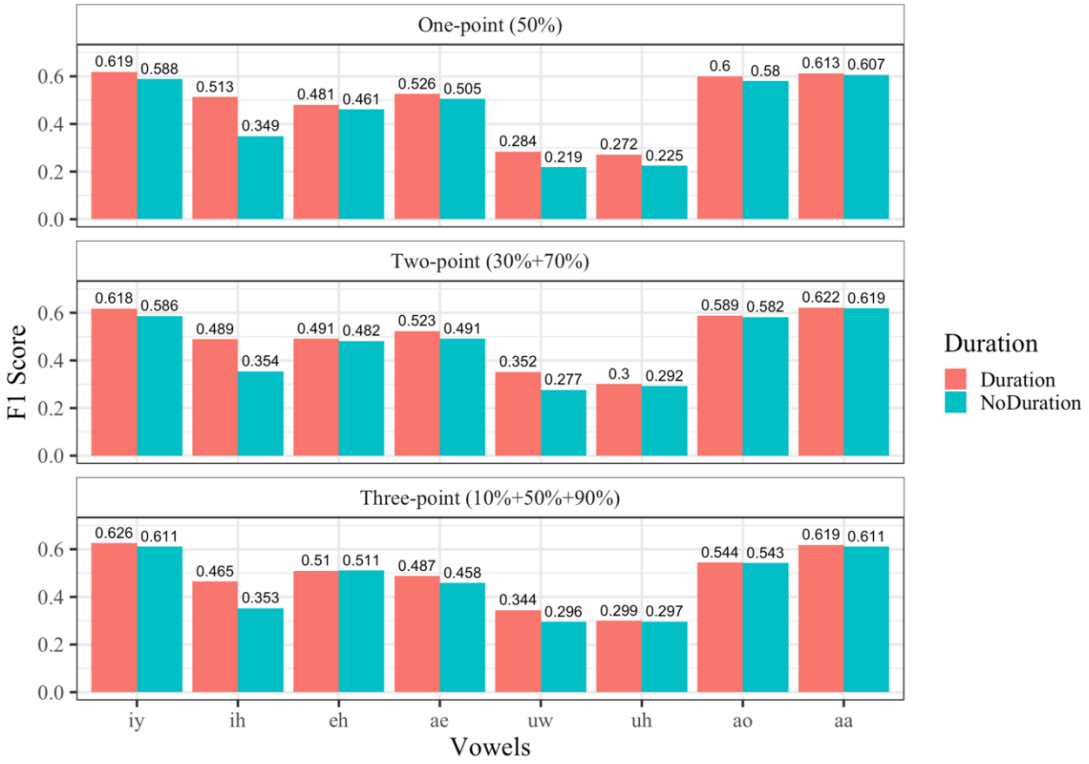

Figure 6. Comparison of F1 Scores with and without Duration

4. Discussion and Conclusion

This paper asked whether the trajectory model classifies vowels better than the steady-state model in natural speech. The results show interesting findings in terms of classification predictors and dynamic cues. All of the models showed the highest accuracy results when the lowest three formants and fundamental frequency were used as predictors. Comparing the classification accuracy with F0, F1, F2, and F3, the trajectory models, the two-point and three-point models, show better performance than the steady-state model, the one-point model, while the two-point and three-point models do not show big differences. In addition, vowel duration facilitates classification accuracy, which is specific to certain vowels. The results align with previous studies, which utilized citation forms (e.g., Hillenbrand et al. 1995) and spontaneous speech (Hong 2021, 2023), indicating that dynamic cues are crucial for vowel classification and further extend the research to natural speech.

While the results contribute to vowel classification studies, several points require further discussion. First, the classification accuracy across all models was generally lower than in previous studies such as Hillenbrand et al. (1995); however, this is an intrinsic result of the data set which is composed of natural speech. In other words, unlike the experimental data, which consists of a limited structure of the syllable, usually the /hVd/ syllable, the data set in this paper includes various consonants before and after the vowels. Moreover, the classification accuracy gets lower with more number of consonants in an experimental setting (e.g., Almurashi et al. 2024). Thus, it is not surprising that conversational data show lower classification accuracy (e.g., Hong 2021, for Korean spontaneous speech). Although the accuracy seems low for natural speech, the accuracy is significantly different from no information rate, which indicates that the results are not random classification. With more constraints on the

consonants before and after the vowels, higher classification is expected even with natural speech data. Another contributing factor to the low accuracy rates may be specific vowels. If we examine the accuracy rates of each vowel, these rates vary among them. Some vowels show relatively higher accuracy rates than others. For example, the vowels 'ih', 'uw', and 'uh' consistently demonstrate lower accuracy rates across all models, whereas the high front vowel 'iy' and the back low vowels 'ao' and 'aa' show higher accuracy rates. The low accuracy rates of 'ih', 'uw', and 'uh' have consequently impacted the overall accuracy. As the accuracy rates for these vowels improve, it is anticipated that the overall accuracy will also increase, which was observed when duration was included as a predictor.

Second, the results show that not all vowels behave the same across the models. Some vowels are affected more by the dynamic information. That is, the vowels 'uw' and 'uh' show better classification in trajectory models, which is illustrated by the confusion matrices and F1 scores (Table 10 and Table 13 compared to Table 7, and Table 14 for F1 scores), which show higher scores for the vowels. Unlike the vowels, some vowels, the low back vowel 'aa', are not affected much by the dynamic cues. For the vowel, there is almost no difference across all models. For some vowels, the steady state is essential for classification. The vowel 'ae' resulted in a higher F1 score in the one-point model than the other two models when vowel duration is not included as a predictor, while the vowels 'ih', 'ae', and 'ao' resulted in higher scores in the one-point model than the other two models when vowel duration is included as a predictor. According to Harrington and Cassidy (1994), vowels require distinct spectral features in classification, depending on whether it is a monophthong or a diphthong. Dynamic cues facilitate the classification of diphthongs, whereas the steady-state feature is more significant for monophthongs. Given that each vowel requires different cues for classification, further research is necessary to investigate the distinguishing characteristics of the vowels 'uw' and 'uh' in comparison to 'ae', focusing on the trajectory patterns.

While this study contributes evidence that trajectory models classify vowels better not only in citation forms but also in natural speech, it does have some limitations. Firstly, this paper examined four sets of predictors: F1, F2; F1, F2, F3; F0, F1, F2; and F0, F1, F2, F3. While the focus is primarily on the set that included F0, F1, F2, and F3, it is important to note that the classification accuracies varied across the other predictor sets. The three-point model using the lowest two formants, F1 and F2, exhibited the lowest accuracy rate, showing only a marginal difference (0.008) compared to the one-point model. Similarly, the model incorporating the lowest three formants, F1, F2, and F3, showed a minor difference of 0.017 from the one-point model. In contrast, when the fundamental frequency, F0, was included, either in the predictor sets with F0, F1, F2 or F0, F1, F2, F3, the three-point model demonstrated the highest accuracy results. In other words, when formants were used as predictors exclusively, the steady-state model performed slightly better than the trajectory models. Conversely, the trajectory model outperformed the steady-state model when F0 was included. As mentioned above, F0 facilitates distinguishing the speaker groups and each vowel, resulting in higher accuracy. In other words, F0 and vowel dynamics interact to classify the vowels.

In addition to F0, the vowels 'ih', 'uw', and 'uh' require further investigation due to their tendency to be confused with one another, which leads to misclassification in all models. The confusion matrices and F1 scores indicate that the vowel 'ih' shows the highest error rate across the models, which is expected given that it is the most frequently used vowel. Nevertheless, the classification tendencies for these vowels suggest that incorporating dynamic cues improves their classification accuracy. That is, the trajectory models classify these vowels better than the steady-state model. Without dynamic cues, that is, within the steady-state model, vowel duration becomes a significant predictor in the classification of these vowels. Specifically, including vowel duration as a predictor substantially reduces the error rate for 'ih'. This indicates that vowel duration is crucial for accurately classifying 'ih' with its neighboring vowels, 'uw' and 'uh'. Particularly in cases where dynamic cues for vowel classification

are lacking, duration serves as an important compensatory predictor. Therefore, in addition to vowel dynamics and F0, duration is an important cue for vowel classification, especially for the vowel 'ih'.

In conclusion, this paper demonstrated that the trajectory models—the two-point and three-point models—outperform the steady-state or one-point model in classifying vowels in natural speech, as evidenced by an analysis of the Buckeye corpus using quadratic discriminant analysis. By exploring the dynamic cues present in natural speech, this study not only reinforces previous research but also sheds light on the classification of each vowel model. Specifically, the paper provides a detailed analysis of each vowel through confusion matrices and F1 scores, enhancing our understanding of which vowels are most influenced by dynamic features and vowel duration. This paper allowed us to identify the vowels that are most frequently misclassified. Furthermore, by clarifying the characteristics of each vowel, this study proposes future avenues for research, suggesting further investigation into specific vowels, and predictors. This study, therefore, makes a novel contribution by examining dynamic cues in a natural speech environment, using machine learning techniques to provide new insights into how vowel classification models can be enhanced beyond traditional citation-based analyses.

References

Adank, P., R. Van Hout and R. Smits. 2004. An acoustic description of the vowels of Northern and Southern Standard Dutch. *The Journal of the Acoustical Society of America* 116(3), 1729-1738.

Almurashi, W., J. Al-Tamimi and G. Khattab. 2020. Static and dynamic cues in vowel production in Hijazi Arabic. *The Journal of the Acoustical Society of America* 147(4), 2917-2927.

Almurashi, W., J. Al-Tamimi and G. Khattab. 2024. Dynamic specification of vowels in Hijazi Arabic. *Phonetica* 81(2), 185-220.

Boersma, P. and D. Weenink. 2023. Praat: Doing phonetics by computer [Computer program].

Harrington, J. and S. Cassidy. 1994. Dynamic and Target Theories of Vowel Classification: Evidence from Monophthongs and Diphthongs in Australian English. *Language and Speech* 37(4), 357-373.

Hillenbrand, J. and R. T. Gayvert. 1993. Vowel Classification Based on Fundamental Frequency and Formant Frequencies. *Journal of Speech, Language, and Hearing Research* 36(4), 694-700.

Hillenbrand, J., L. A. Getty, M. J. Clark and K. Wheeler. 1995. Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America* 97(5), 3099-3111.

Hillenbrand, J. 2013. Static and Dynamic Approaches to Vowel Perception. In G. S. Morrison and P. F. Assmann, eds., *Vowel Inherent Spectral Change*, 9-30. Springer.

Hillenbrand, J., M. J. Clark and T. M. Nearey. 2001. Effects of consonant environment on vowel formant patterns. *The Journal of the Acoustical Society of America* 109(2), 748-763.

Hong, S. 2021. Roles of temporal patterns of vowel-intrinsic cues in model identification of Korean vowels in spontaneous speech. *Studies in Phonetics, Phonology, and Morphology* 27(2), 321-351.

Hong, S. 2023. Roles of dynamic patterns of lower formants, vowel identity, and gender in predicting postvocalic consonant place in Korean spontaneous speech. *Studies in Phonetics, Phonology, and Morphology* 29(2), 211-246.

Kelleher, J. D., B. Mac Namee and A. D'Arcy. 2020. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies* (second edition). The MIT Press.

Kuhn, M. 2008. Building predictive models in R using the caret package. *Journal of Statistical Software* 28(5), 1-26.

Morrison, G. S. 2013. Theories of Vowel Inherent Spectral Change. In G. S. Morrison and P. F. Assmann, eds., *Vowel Inherent Spectral Change*, 31-48. Springer.

Nearey, T. M. and P. F. Assmann. 1986. Modeling the role of inherent spectral change in vowel identification. *The Journal of the Acoustical Society of America* 80(5), 1297-1308.

Neel, A. T. 2004. Formant detail needed for vowel identification. *Acoustics Research Letters Online* 5(4), 125-131.

Peterson, G. E. and H. L. Barney. 1952. Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America* 24(2), 175-184.

Pitt, M.A., L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume and E. Fosler-Lussier. 2007. Buckeye corpus of conversational speech (2nd release). [www.buckeyecorpus.osu.edu].

Pitt, M A., K. Johnson, E. Hume, S. Kiesling and W. Raymond. 2005. The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication* 45(1), 89-95.

R Core Team. 2023. R: A language and environment for statistical computing. [www.R-project.org].

Watson, C. I. and J. Harrington. 1999. Acoustic evidence for dynamic formant trajectories in Australian English vowels. *The Journal of the Acoustical Society of America* 106(1), 458-468.

Whalen, D. H., and A. G. Levitt. 1995. The universality of intrinsic F0 of vowels. *Journal of Phonetics* 23(3), 349-366.

Yoon, K. 2021. *Praat & Scripting*. Book Kyul.

Zahorian, S. A. and A. J. Jagharghi. 1993. Spectral-shape features versus formants as acoustic correlates for vowels. *The Journal of the Acoustical Society of America* 94(4), 1966-1982.

Examples in: English
Applicable Languages: English
Applicable Level: Primary/Secondary/Tertiary

Acknowledgement
I would like to thank two anonymous reviewers for their constructive feedback. Any remaining shortcomings are the author's responsibility.