Vowel Classification in Conversational Speech Corpus

Hyun Jin Hwangbo

Pukyong National University

Introduction and Background. Since the seminal work of Hillenbrand et al. (1995), there has been widespread acknowledgment of the dynamic nature of vowels, emphasizing the significance of their onset and offset for accurate identification. Hillenbrand et al.'s investigation showed that incorporating the onset and offset of vowels alongside the steady state in quadratic discriminant analysis (QDA) led to enhanced accuracy rates in vowel categorization. Specifically, their findings revealed a heightened accuracy rate when formant values (F1, F2, F3) at 20% and 80% of the vowel duration were considered, compared to when values solely at 50% were utilized. To establish the generalizability of these findings within natural conversational speech, this study examines the Buckeye Corpus of conversational speech (Pitt et al., 2007). The primary objective of this research is to classify vowels through the implementation of the machine learning technique, QDA, with a training dataset, and then test the model with a new dataset. While the results of this study align with the trends observed by Hillenbrand et al. (1995), the accuracy rates pertaining to the test dataset requires further examination and discussion. **Data Analysis.** The Buckeye Corpus comprises spontaneous speech of a total of 40 adults. Utilizing Praat (Boersma and Weenink, 2023) in conjunction with a modified version of the Praat script based on Yoon (2021), formants and fundamental frequency (F0) were extracted. Formants were sampled at 10% intervals throughout the vowel duration, encompassing nine monophthongs for analysis: [i, I, E, æ, A, O, U, u, @/2].[1] Data points with undefined values or errors during formant extraction or F0 determination were systematically excluded. The dataset was partitioned into a training set (70% of the data) and a test set (30% of the data) to facilitate classification accuracy assessment. To address dataset imbalance during the training phase, vowel occurrences were upsampled to 13,535 instances, resulting in a total of 121,815 occurrences across all vowels. QDA with a cross-validation (k = 10) approach was conducted using R (R Core Team, 2023) and the 'caret' package (Kuhn, 2008). Three predictor models were explored, incorporating various combinations of F1, F2, F3 and F0. These models included a one-predictor model utilizing formants at 50% of the vowel duration, a two-predictor model incorporating formants at 20% and 80% duration, and a three-predictor model integrating formants at 20%, 50%, and 80% duration. Additionally, vowel duration was included as a predictor in all models. **Results.** Table 2 presents the accuracy rates of vowel classification derived from the training set.[2] As the table shows, accuracy rates exhibit an upward trend with an increasing number of predictors and greater utilization of formant values and F0. Notably, the two-predictor model featuring duration alongside all formant values and F0 achieves the highest accuracy rate. This observed trend aligns with the findings of Hillenbrand et al. Table 3 depicts the accuracy rates of vowel classification within the test set. Interestingly, the one-predictor model incorporating duration alongside all formant values and F0 demonstrates the highest accuracy rate, followed by the two-predictor model with duration. This divergence from the training set results suggests a nuanced classification pattern. **Conclusion and Discussion.** The outcomes of QDA underscore the significance of onset and offset predictors alongside comprehensive formant information and duration. While the observed trends in learning out-comes mirror those reported in previous research, disparities emerge upon analysis of the test dataset. Notably, the one-predictor model featuring duration exhibits superior accuracy com-pared to the two-predictor model. It appears that formant values and duration play a pivotal role in vowel classification during the steady state. This discrepancy in classification accuracy between the training and test datasets warrants careful consideration. The observed discrepancy in test accuracy may arise from various factors including algorithmic consideration such as overfitting to the training data, differences in sample sizes, and other linguistic necessitating further investigation for model refinement to enhance classification performance.

	i	ɪ	ɛ	æ	ɑ	ɔ	ʊ	u	ə/ʌ	Total
Training	5075	13535	9144	3681	3731	2528	1722	1806	12179	53401
Test	2174	5800	3918	1577	1599	1083	737	773	5219	22880
Total	7249	19335	13062	5258	5330	3611	2459	2579	17398	

Table 1: Number of occurrences of each vowel in training and test set

	One-predictor (50%)		Two-predictor (20%, 80%)		Three-predictor (20%, 50%, 80%)	
	no dur	dur	no dur	dur	no dur	dur
F1, F2	0.411	0.439	0.421	0.448	0.431	0.448
F1, F2, F3	0.432	0.462	0.455	0.479	0.454	0.468
F0, F1, F2	0.436	0.459	0.449	0.470	0.454	0.459
F0, F1, F2, F3	0.449	0.476	0.469	0.490	0.463	0.474

Table 2: Accuracy results of vowel classification using quadratic discriminant analysis (QDA)

	One-predictor (50%)		Two-predictor (20%, 80%)		Three-predictor (20%, 50%, 80%)	
	no dur	dur	no dur	dur	no dur	dur
F1, F2	0.33	0.407	0.309	0.369	0.307	0.338
F1, F2, F3	0.355	0.426	0.339	0.387	0.331	0.356
F0, F1, F2	0.36	0.438	0.352	0.415	0.378	0.41
F0, F1, F2, F3	0.373	0.441	0.365	0.414	0.38	0.408

Table 3: Accuracy results of vowel classification with QDA of test set

References

Boersma, P. and Weenink, D. (2023). Praat: doing phonetics by computer [Computer program]. Version 6.4.02, retrieved 30 December 2023 from https://www.praat.org.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic Characteristics of American English Vowels. *Journal of the Acoustical Society of America*, 97:3099–3111.

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

Pitt, M., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., and Fosler-Lussier, E. (2007). Buckeye corpus of conversational speech. [www.buckeyecorpus.osu.edu]. Columbus, OH: Department of Psychology, Ohio State University (Distributor).

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Yoon, K. (2021). *Praat & Scripting*. Book Kyul, Seoul.

[1] The corpus does not distinguish [ə] and [ʌ], and thus, the two vowels were analyzed as a single vowel.

[2] In Table 2 and 3, 'no dur' indicates the analysis without duration and 'dur' indicates the analysis including vowel duration as a predictor