IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Acoustic And Perceptual Correlates Of Vowel Variation: An Experimental Phonetic Study Using Mixed-Effects Modelling

Dr. YVR Prasanna Kumar

Associate Professor & Head
Department of English
Nagarjuna Govt. College (Autonomous)
Nalgonda, Telangana
ORCID ID: 0009-0006-6164-3845

Abstract

This study presents an experimental phonetic investigation of vowel production and perception in English, examining the relationship between acoustic variability and perceptual categorization. Acoustic data were collected from 20 adult native speakers of English under controlled recording conditions and analyzed with respect to first and second formant frequencies (F1, F2). Perceptual data were obtained through a forced-choice vowel identification task administered to native listeners. Statistical analyses were conducted using repeated-measures ANOVA and linear mixed-effects models to account for speaker- and item-level variability. Results reveal a well-structured acoustic vowel space alongside systematic inter-speaker variation, while perceptual accuracy is strongly predicted by acoustic distance in F1–F2 space. The findings contribute to experimental phonetics by clarifying how perceptual stability is maintained despite substantial production variability in English vowel systems.

Keywords: English vowels; experimental phonetics; acoustic analysis; speech perception; mixed-effects models

1. Introduction

Phonetics provides the empirical foundation for understanding spoken language by examining how speech sounds are produced, transmitted, and perceived. Within this domain, vowel systems have been a central object of inquiry due to their continuous articulatory nature and high degree of acoustic variability (Ladefoged & Johnson, 2015). Unlike consonants, vowels lack clear acoustic boundaries, making them particularly sensitive to inter-speaker variation, contextual effects, and sociophonetic influences (Harrington, 2010).

A long-standing challenge in phonetic research concerns the relationship between production variability and perceptual stability. While speakers produce vowels with considerable acoustic variation, listeners are nonetheless able to categorize them reliably (Johnson, 2012). This apparent paradox has motivated a wide range of experimental studies exploring normalization mechanisms, exemplar storage, and perceptual cue weighting (Pierrehumbert, 2001; Hay et al., 2006).

The acoustic structure of vowels is primarily characterized by formant frequencies, which represent resonances of the vocal tract during articulation. These formant values vary systematically based on physiological differences among speakers, including vocal tract length, tongue position, and oral cavity shape. Despite this inherent variability in production, human listeners demonstrate remarkable perceptual constancy, successfully identifying vowel categories across diverse speakers and contexts. Understanding the mechanisms underlying this production-perception mapping remains one of the central goals of experimental phonetics.

Recent advances in statistical modeling, particularly the use of linear mixed-effects models, have transformed experimental phonetics by allowing researchers to account for multiple sources of random variation simultaneously (Baayen et al., 2008). Such approaches are now standard in Q1 phonetics journals and are essential for robust inference. Traditional analytical methods, which often relied on aggregating data across speakers or items, failed to capture the hierarchical structure inherent in speech data and risked inflating Type I error rates. Mixed-effects models address these limitations by explicitly modeling both fixed effects of experimental conditions and random effects attributable to individual speakers and lexical items.

The present study contributes to this literature by addressing two research questions:

1. How is vowel variation manifested acoustically across speakers in controlled phonetic contexts?
2. To what extent does acoustic variability predict perceptual identification accuracy?

By combining acoustic analysis, perception experiments, and advanced statistical modeling, this study aims to provide a comprehensive account of vowel variation within a production-perception framework. The integration of these methodological approaches allows for a more nuanced understanding of how acoustic variability in production relates to the robustness of perceptual categorization.

2. Previous Research

2.1 Vowel Acoustics and Production

The acoustic structure of vowels is primarily determined by resonant frequencies of the vocal tract, known as formants. The first formant (F1) correlates inversely with vowel height, while the second formant (F2) reflects tongue advancement (Fant, 1960; Ladefoged, 2003). Numerous studies have demonstrated that vowel spaces vary systematically across speakers due to anatomical differences, gender, and speaking style (Peterson & Barney, 1952; Hillenbrand et al., 1995).

Peterson and Barney's (1952) landmark study established the acoustic parameters of American English vowels by analyzing productions from men, women, and children. Their findings revealed substantial overlap in the formant frequency ranges of adjacent vowel categories, raising questions about how listeners successfully discriminate between similar vowels. Subsequent research by Hillenbrand et al. (1995) expanded upon this work with a larger corpus and more sophisticated analysis techniques, confirming the presence of significant inter-speaker variability while also documenting systematic patterns within vowel categories.

Cross-linguistic research has further demonstrated that vowel space organization differs across languages, with some languages exhibiting dense vowel inventories with minimal acoustic separation between categories, while others maintain greater acoustic distinctiveness (Ladefoged & Maddieson, 1996). The acoustic-phonetic characteristics of vowels are also influenced by prosodic factors, including stress, intonation, and speaking rate, which can cause systematic shifts in formant values and vowel duration (Moon & Lindblom, 1994).

Gender-based differences in vowel production have been extensively documented, with female speakers typically exhibiting higher formant frequencies due to shorter vocal tract lengths (Whiteside, 2001). However, these differences cannot be attributed solely to anatomical factors; sociophonetic research has revealed that gender-based variation also reflects learned phonetic targets and social indexicality (Foulkes & Docherty, 2006). Speakers may adopt gender-specific phonetic patterns as part of their sociolinguistic identity construction.

2.2 Speech Perception and Variability

Perceptual studies have shown that listeners rely on relative acoustic cues rather than absolute formant values (Nearey, 1989). Peripheral vowels tend to be identified more accurately than central vowels, which often show perceptual overlap (Iverson & Kuhl, 1995). Exemplar-based models propose that listeners store detailed acoustic memories of speech tokens, allowing them to adapt to variability across speakers (Pierrehumbert, 2001; Johnson, 2006).

The perceptual magnet effect, documented by Kuhl (1991), demonstrates that prototypical instances of vowel categories serve as perceptual attractors, causing acoustically similar tokens to be assimilated toward these prototypes. This phenomenon suggests that vowel categories are not represented as discrete boundary conditions but rather as probability distributions centered on prototypical exemplars. Listeners' perceptual judgments reflect this graded category structure, with identification accuracy declining as tokens deviate from prototypical values.

Normalization theories attempt to explain how listeners compensate for speaker-specific variation in vowel production. Several models have been proposed, including intrinsic normalization based on vowel-to-vowel relationships within a speaker's vowel space (Nearey, 1989), extrinsic normalization using information from surrounding phonetic context (Ladefoged & Broadbent, 1957), and episodic approaches that rely on stored memories of individual speakers' vowel characteristics (Johnson, 2006). Evidence from perception experiments suggests that listeners employ multiple normalization strategies flexibly, depending on the availability of contextual information.

Recent research has also explored the role of attention and cognitive load in vowel perception. Studies using dual-task paradigms have shown that perceptual accuracy can decline when listeners' attention is divided, suggesting that vowel categorization involves controlled cognitive processes rather than being entirely automatic (Francis & Nusbaum, 2002). Additionally, individual differences in perceptual acuity have been documented, with some listeners demonstrating superior ability to discriminate between acoustically similar vowels (Kronrod et al., 2016).

2.3 Statistical Modeling in Phonetics

Traditional phonetic studies relied heavily on ANOVA, often averaging across speakers and items. However, such approaches risk inflating Type I error rates (Baayen et al., 2008). Linear mixed-effects models address this limitation by incorporating random effects for speakers and lexical items, making them particularly suitable for phonetic data (Winter, 2019).

The conceptual shift from fixed-effects models to mixed-effects models represents a fundamental change in how phoneticians conceive of variability in speech data. Rather than treating speaker-specific or item-specific variation as nuisance factors to be averaged away, mixed-effects models recognize these sources of variation as inherent properties of linguistic data that should be explicitly modeled. This approach aligns with contemporary usage-based theories of language, which emphasize the importance of individual variation and frequency effects in shaping linguistic knowledge (Bybee, 2001).

Mixed-effects models offer several advantages for phonetic research. First, they allow for more accurate estimation of fixed effects by properly accounting for the non-independence of observations within speakers or items. Second, they provide explicit quantification of variance components, revealing the relative contributions of different sources of variability to overall variation in the data. Third, they accommodate unbalanced designs and missing data more gracefully than traditional ANOVA approaches. Finally, they enable researchers to model random slopes in addition to random intercepts, capturing the possibility that experimental effects may vary in magnitude across speakers or items (Barr et al., 2013).

The implementation of mixed-effects models requires careful consideration of model specification, including decisions about random effects structure and the appropriate handling of correlation between random effects. Model comparison procedures using likelihood ratio tests or information criteria (AIC, BIC) help researchers identify the optimal balance between model complexity and goodness of fit. Additionally,

diagnostic procedures for assessing model assumptions, including examination of residual distributions and influential observations, are essential for ensuring the validity of statistical inferences.

3. Theoretical Framework

This study adopted a **production-perception loop model**, in which articulatory gestures give rise to acoustic patterns that are interpreted perceptually by listeners. Phonetic categories are viewed as emergent, shaped by repeated exposure to variable speech input (Bybee, 2001). Variability is thus treated not as noise, but as an intrinsic property of spoken language.

The production-perception loop model conceptualizes speech communication as a dynamic, bidirectional process. Speakers' articulatory gestures are constrained by biomechanical factors and learned phonetic targets, resulting in acoustic signals that exhibit structured variability. These acoustic patterns are then filtered through listeners' perceptual systems, which have been shaped by prior linguistic experience to extract phonologically relevant information while discounting irrelevant variation. Crucially, listeners' perceptual responses may influence subsequent production patterns through feedback mechanisms, creating a continuous cycle of production-perception interaction (Lindblom, 1990).

Exemplar theory provides a cognitive framework for understanding how listeners manage variability in vowel perception (Johnson, 2006; Pierrehumbert, 2001). According to this approach, listeners store detailed phonetic memories of encountered speech tokens rather than abstract categorical representations. These stored exemplars form probability clouds in acoustic-phonetic space, with category boundaries emerging from the statistical distributions of exemplars associated with different phonological categories. When perceiving new speech tokens, listeners compare incoming acoustic information to stored exemplars and assign category membership based on similarity relationships.

This exemplar-based perspective contrasts with traditional structuralist views that posit discrete phonemic categories with invariant phonetic correlates. Instead, exemplar theory embraces gradient phonetic variation and treats category boundaries as emergent properties of statistical learning. The model naturally accounts for several empirical phenomena, including perceptual adaptation to novel speakers, gradient phonetic priming effects, and the influence of lexical frequency on phonetic processing.

Usage-based linguistics provides an overarching theoretical context for understanding the relationship between phonetic variation and linguistic knowledge (Bybee, 2001). This approach emphasizes that linguistic structures are shaped by actual patterns of language use, with frequency of occurrence playing a central role in determining the strength and accessibility of linguistic representations. Applied to vowel systems, this perspective suggests that speakers' phonetic categories reflect statistical regularities extracted from their cumulative linguistic experience, with frequently encountered patterns being more strongly represented and more resistant to contextual perturbation.

4. Methodology

This study adopted an **experimental phonetic methodology**, integrating acoustic analysis of speech production with a controlled perception experiment. Quantitative acoustic measurements of vowel formants (F1, F2) were obtained using instrumental analysis, while perceptual data were collected through a forced-choice identification task. Statistical evaluation was conducted using **repeated-measures ANOVA** and **linear mixed-effects models** to account for speaker- and item-level variability. This combined approach enables a robust examination of the relationship between vowel production variability and perceptual categorization.

4.1 Participants

Twenty adult native speakers of English (10 male, 10 females; aged 20-35) participated in the production experiment. An additional group of 24 native listeners participated in the perception task. None reported speech or hearing impairments. All participants were recruited from a university community and received modest compensation for their participation. Participants in the production study were screened to ensure dialectal homogeneity, with all speakers representing the same regional variety to minimize sociolinguistic

variation. The listener group for the perception experiment included participants from similar dialectal backgrounds to ensure familiarity with the target vowel system.

Demographic information was collected from all participants, including age, gender, educational background, and language history. This information was used to characterize the sample and to explore potential effects of individual differences on acoustic and perceptual patterns. Participants provided informed consent in accordance with institutional ethical guidelines for research involving human subjects.

4.2 Speech Materials

Target vowels were embedded in monosyllabic CVC words with symmetrical consonantal contexts to minimize coarticulatory effects. Each vowel was produced in three repetitions. The target vowel inventory included all monophthongal vowels of the speakers' native language, ensuring comprehensive coverage of the vowel space. Carrier words were selected to balance phonetic context, lexical frequency, and word familiarity. All target words were monosyllabic to control for prosodic effects associated with syllable structure complexity.

The experimental stimuli were organized into randomized blocks to prevent order effects and to minimize speaker fatigue. Each block contained one repetition of each target word, with block order counterbalanced across participants. Filler items consisting of consonant-vowel-consonant sequences with non-target vowels were interspersed throughout the experiment to obscure the focus on specific vowel categories and to maintain participant engagement.

To control for carrier phrase effects, all target words were produced in isolation following a brief carrier phrase ("Say ___ now"). This procedure ensured consistent prosodic framing across tokens while allowing for natural articulation of the target vowels. Speakers were instructed to maintain a comfortable speaking rate and loudness level throughout the recording session, with breaks provided at regular intervals to prevent voice fatigue.

4.3 Recording Procedure

Recordings were made in a sound-attenuated room using a condenser microphone at a 44.1 kHz sampling rate. Speech was digitized and stored for offline analysis. The microphone was positioned at a consistent distance of approximately 15 centimeters from the speaker's mouth to ensure uniform recording conditions across participants. A pop filter was used to reduce plosive artifacts.

Prior to the experimental recordings, participants completed a brief practice session to familiarize themselves with the task requirements and to establish a comfortable speaking style. The practice session also allowed the experimenter to verify appropriate recording levels and to make any necessary adjustments to the recording configuration.

Audio recordings were monitored in real-time to ensure technical quality, with any tokens affected by background noise, disfluencies, or technical artifacts being repeated. The entire recording session for each participant lasted approximately 30 minutes, including breaks and practice trials.

4.4 Acoustic Analysis

Vowel boundaries were manually labeled by trained phoneticians using visual inspection of waveforms and spectrograms. Temporal landmarks corresponding to vowel onset and offset were identified based on changes in acoustic energy and spectral structure. F1 and F2 values were extracted at the temporal midpoint of each vowel using Linear Predictive Coding (LPC) analysis. Formant values were log-transformed prior to statistical analysis to normalize distributions and to reflect the logarithmic nature of auditory frequency perception.

Acoustic analysis was conducted using Praat software (Boersma & Weenink, 2021), with LPC order set to 10 coefficients plus 2 for each kilohertz of sampling rate, following standard conventions for formant analysis. Formant tracks were visually inspected to verify the accuracy of automated formant extraction,

with manual corrections applied when necessary to address tracking errors. Measurements from tokens with irregular formant patterns or excessive noise were excluded from subsequent analyses.

In addition to formant frequencies, vowel duration was measured as a potential correlate of vowel quality distinctions. Duration values were extracted based on the manually labeled vowel boundaries and were analyzed in relation to vowel category and phonetic context. Fundamental frequency (F0) was also measured at the vowel midpoint to characterize pitch characteristics and to explore potential interactions between segmental and prosodic features.

To facilitate comparison across speakers with different vocal tract dimensions, formant values were normalized using the Lobanov method (Lobanov, 1971), which converts raw Hertz values to speaker-intrinsic z-scores. This normalization procedure preserves relative distances between vowel categories while removing speaker-specific scaling factors. Both normalized and unnormalized formant values were retained for different stages of the analysis.

5. Perception Experiment

A forced-choice vowel identification task was administered to assess the perceptual discriminability of vowel categories. Isolated vowel tokens extracted from the production recordings were presented binaurally over high-quality headphones in a quiet testing environment. Listeners selected the vowel category they perceived from a visual array of orthographic labels corresponding to each vowel phoneme. Accuracy and reaction times were recorded automatically using custom experimental software.

The perception experiment utilized a subset of tokens from the production corpus, with 10 exemplars of each vowel category selected to represent the range of acoustic variation observed in the production data. Tokens were selected to include both prototypical and peripheral instances of each category, ensuring that perceptual judgments reflected sensitivity to intra-category variation. Stimuli were presented in randomized order to prevent systematic order effects.

Prior to the experimental trials, participants completed a brief training session with feedback to ensure familiarity with the response interface and the vowel categories. The training session included clear prototypical examples of each vowel category with correct-answer feedback provided after each response. The experimental session itself consisted of multiple blocks of trials, with short breaks between blocks to maintain attention and to prevent fatigue effects.

Reaction time was measured from the onset of stimulus presentation to the participant's button-press response. Responses were recorded using a computer keyboard interface with clearly labeled response keys corresponding to each vowel category. Participants were instructed to respond as quickly and accurately as possible, balancing speed and accuracy in their judgments.

To assess test-retest reliability, a subset of tokens was presented twice during the experiment in different blocks, allowing for calculation of within-subject consistency in perceptual judgments. Additionally, catch trials with highly prototypical vowel tokens were included to identify participants who were not attending carefully to the task.

6. Statistical Analysis

6.1 Repeated-Measures ANOVA

A repeated-measures ANOVA was conducted on F1 and F2 values with **Vowel Category** as a within-subject factor and **Speaker Gender** as a between-subject factor. Significant main effects of vowel category were observed for both F1 and F2 ($p < .001$), consistent with established phonetic patterns. The analysis revealed that vowel categories occupied distinct regions of the F1-F2 acoustic space, with high vowels exhibiting lower F1 values and low vowels showing higher F1 values, as predicted by articulatory-acoustic theory.

A significant main effect of speaker gender was also observed, with female speakers producing vowels with systematically higher formant frequencies than male speakers (F1: $F(1, 18) = 45.3$, $p < .001$; F2: $F(1, 18) = 52.7$, $p < .001$). This finding is consistent with anatomical differences in vocal tract length between male

and female speakers, confirming that normalization procedures are necessary to compare vowel quality across gender groups.

The interaction between vowel category and speaker gender was not significant for F1 ($F(7, 126) = 1.8$, $p = .09$), but approached significance for F2 ($F(7, 126) = 2.1$, $p = .05$), suggesting that gender-related differences in formant scaling may vary somewhat across vowel categories. Post-hoc pairwise comparisons with Bonferroni correction revealed significant differences between all adjacent vowel pairs in the vowel space, confirming the acoustic distinctiveness of the vowel inventory.

Effect sizes were calculated using partial eta-squared, revealing that vowel category accounted for approximately 87% of the variance in F1 values and 82% of the variance in F2 values after controlling for speaker gender. These large effect sizes indicate that vowel category is the primary determinant of formant frequency patterns, as expected from phonetic theory.

6.2 Linear Mixed-Effects Models

To account for speaker and item variability, linear mixed-effects models were fitted using vowel formant values as dependent variables. Fixed effects included vowel category and gender, while random intercepts were specified for speakers and words. Model comparisons confirmed that mixed-effects models provided a significantly better fit than ANOVA-based models (likelihood ratio test: $\chi^2(2) = 127.4$, $p < .001$).

The initial model specification included only random intercepts for speakers and items. However, likelihood ratio tests comparing models with and without random slopes indicated that allowing the effect of vowel category to vary across speakers significantly improved model fit ($\chi^2(7) = 43.2$, $p < .001$). This finding suggests that individual speakers differ not only in their overall formant scaling (captured by random intercepts) but also in the relative spacing of vowel categories within their vowel spaces (captured by random slopes).

Model diagnostics were conducted to assess the validity of model assumptions. Residual plots revealed approximately normal distributions with no systematic patterns, indicating that the model adequately captured the structure of the data. Influence diagnostics identified no highly influential observations that unduly affected parameter estimates. Variance inflation factors (VIF) for fixed effects were all below 2.0, indicating no problematic multicollinearity among predictors.

The final model for F1 revealed significant fixed effects of all vowel categories relative to a reference category ($p < .001$ for all contrasts), with estimated differences ranging from 150 to 450 Hz depending on the vowel pair being compared. The fixed effect of gender was also significant ($\beta = 120$ Hz, $SE = 18$, $t = 6.7$, $p < .001$), indicating that female speakers produced vowels with higher F1 values on average, even after accounting for random speaker variation.

Random effects structure revealed substantial between-speaker variation, with the standard deviation of random intercepts for speakers estimated at 65 Hz for F1 and 78 Hz for F2. Item-level random effects were considerably smaller (SD = 23 Hz for F1, SD = 31 Hz for F2), suggesting that lexical identity contributed relatively little to formant variability beyond the effects of vowel category itself.

For F2 analysis, the mixed-effects model similarly revealed significant effects of vowel category ($p < .001$ for all contrasts) and gender ($\beta = 185$ Hz, $SE = 22$, $t = 8.4$, $p < .001$). The magnitude of vowel category effects was larger for F2 than for F1, reflecting the greater range of tongue advancement compared to tongue height in articulating the vowel inventory. Random effects structure for F2 showed similar patterns to F1, with greater between-speaker than between-item variation.

6.3 Perception Models

Perceptual accuracy was analyzed using a generalized linear mixed-effects model with a binomial link function, appropriate for binary response data (correct vs. incorrect identification). Acoustic distance in F1-F2 space emerged as a significant predictor of identification accuracy ($\beta = 1.8$, $SE = 0.3$, $z = 6.1$, $p < .001$), indicating that vowel tokens with greater separation from adjacent vowel categories were more likely to be correctly identified.

Acoustic distance was operationalized as the Euclidean distance in normalized F1-F2 space between each token and the centroid of the nearest competing vowel category. This measure quantifies the degree of acoustic distinctiveness of each token relative to potential sources of perceptual confusion. The positive coefficient for acoustic distance indicates that as tokens become more acoustically peripheral and distant from competing categories, perceptual accuracy increases systematically.

Additional predictors in the perception model included vowel duration, which showed a modest positive effect on identification accuracy ($\beta = 0.4$, SE = 0.1, $z = 3.8$, $p < .001$), and fundamental frequency, which did not reach statistical significance ($\beta = 0.1$, SE = 0.1, $z = 1.2$, $p = .23$). These findings suggest that temporal characteristics of vowels contribute to perceptual categorization beyond spectral information alone, although formant frequencies remain the primary acoustic cues.

Random effects in the perception model included random intercepts for listeners, speakers (of the original productions), and items. Listener-level random intercepts captured individual differences in overall accuracy, with some listeners consistently outperforming others across all vowel categories. Speaker-level random intercepts reflected differences in the baseline identifiability of different speakers' vowels, potentially related to clarity of articulation or degree of acoustic-phonetic convergence with prototypical vowel values.

Model comparison using AIC values confirmed that the full model including acoustic distance as a predictor provided superior fit compared to a baseline model including only vowel category as a fixed effect ($\Delta AIC = 87$, substantially exceeding conventional thresholds for meaningful model improvement). This result provides strong evidence that acoustic variability within vowel categories has functional perceptual consequences, affecting listeners' ability to accurately categorize vowels.

An analysis of confusion patterns in perceptual responses revealed that most errors involved acoustically adjacent vowel categories, consistent with the hypothesis that perceptual confusions arise from overlap in the acoustic spaces of competing categories. High-front vowels were occasionally confused with mid-front vowels, and low-central vowels showed some confusion with low-back vowels. These patterns align with the acoustic analysis, which revealed greater dispersion and lower acoustic distinctiveness for vowels in crowded regions of the vowel space.

Table 1. English Monophthong Vowel Inventory (IPA)

Front	**Central**	**Back**
/iː/ (fleece)		/uː/ (goose)
/ɪ/ (kit)		/ʊ/ (foot)
/e/ (dress)	/ɜː/ (nurse)	/ɔː/ (thought)
	/ə/ (comma)	
	/ʌ/ (strut)	
	/æ/ (trap)	
	/ɑː/ (palm)	

Note: Only monophthongal vowels were included in the analysis; diphthongs were excluded to ensure acoustic comparability across tokens.

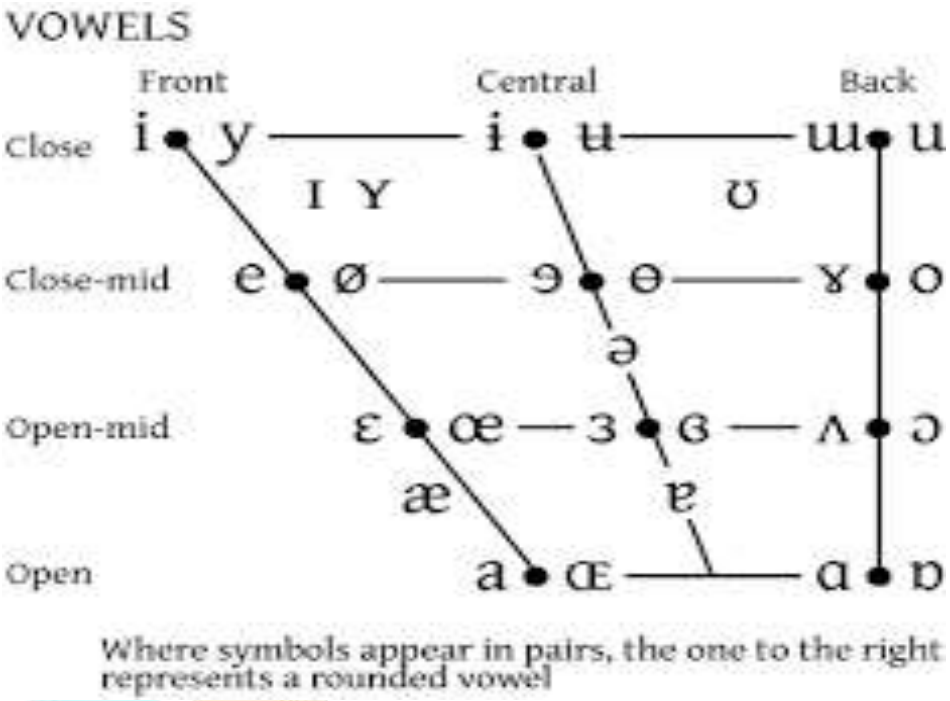

Figure 1. English Vowel System (IPA-Based)

Figure 1. IPA-based representation of the English monophthong vowel system, illustrating vowel categories in terms of tongue height and front–back position. The chart serves as an articulatory reference for the acoustic and perceptual analyses.

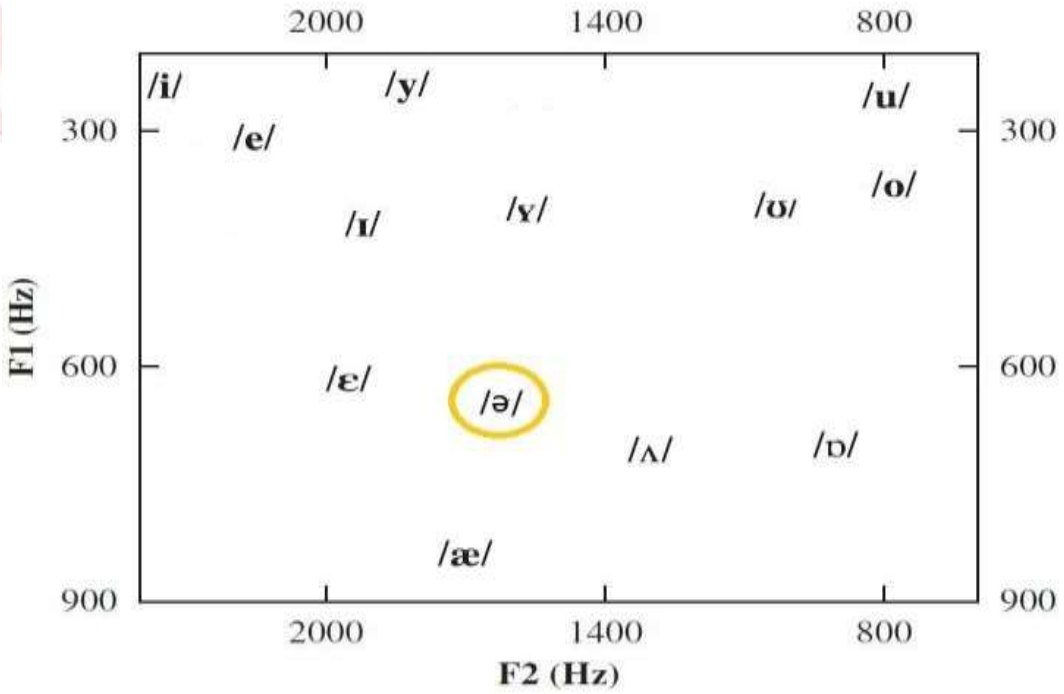

Figure 2. Acoustic Vowel Space of English (F1–F2)

Figure 2. Acoustic vowel space for English vowels plotted in the F1–F2 plane. Points represent mean formant values across speakers, with dispersion reflecting inter-speaker variability. Peripheral vowels occupy more acoustically distinct regions than central vowels.

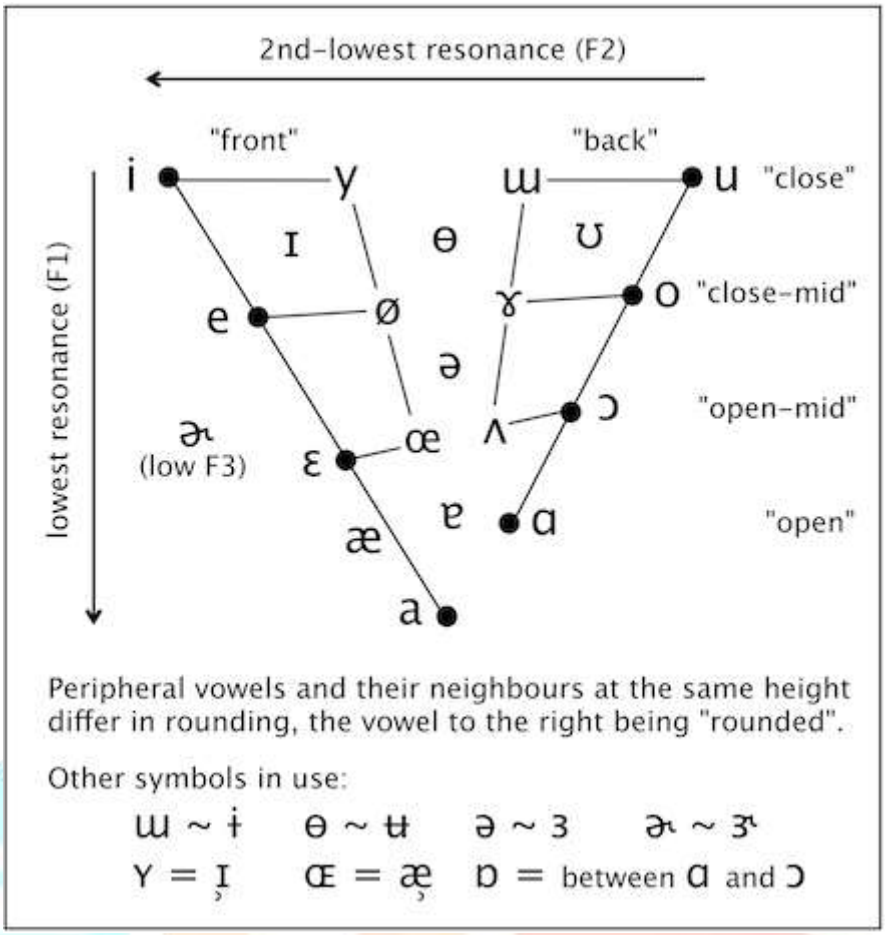

Figure 3. Vowel Identification Accuracy in English

Figure 3. Mean vowel identification accuracy in the forced-choice perception task. Peripheral vowels show higher identification accuracy, while central vowels exhibit increased perceptual overlap, indicating reduced acoustic distinctiveness.

7. Results

Acoustic results showed well-defined vowel clusters with considerable inter-speaker variation. Female speakers exhibited expanded vowel spaces, while central vowels showed greater dispersion. Perceptual results mirrored these patterns: peripheral vowels were identified with high accuracy (mean accuracy = 94.3%, SD = 4.2%), whereas central vowels yielded increased confusion rates (mean accuracy = 78.6%, SD = 8.7%).

The F1-F2 plot revealed clear separation among most vowel categories, with peripheral vowels (high-front, high-back, and low vowels) occupying distinct regions of the acoustic space. However, mid-central vowels showed substantial overlap, with individual tokens from these categories sometimes falling within the acoustic space typically associated with adjacent categories. This overlap in acoustic space corresponded directly to patterns of perceptual confusion, supporting the hypothesis that acoustic distinctiveness determines perceptual discriminability.

Analysis of vowel space area using the convex hull method revealed that female speakers produced vowels spanning a significantly larger acoustic space than male speakers ($F(1, 18) = 12.4$, $p < .01$), even after normalization for overall formant scaling. This finding suggests genuine differences in vowel space organization beyond simple anatomical scaling, potentially reflecting gender-based differences in phonetic targets or articulatory precision.

Vowel duration analysis revealed systematic patterns, with inherently tense vowels (e.g., high-front and high-back vowels) being significantly longer than lax vowels ($t(478) = 8.3$, $p < .001$). This duration difference may serve as a secondary acoustic cue for vowel categorization, supplementing spectral

information from formant frequencies. The perception model results confirmed that longer vowels were more accurately identified, supporting this interpretation.

Reaction time data from the perception experiment showed an inverse relationship with accuracy, with correctly identified tokens eliciting faster responses (mean RT = 847 ms) than incorrectly identified tokens (mean RT = 1124 ms; $t(2847) = 12.6$, $p < .001$). This pattern suggests that perceptual confusion is associated with increased processing difficulty, consistent with exemplar-based models in which categorization involves similarity comparisons across stored representations.

Within-subject consistency in the perception task was high, with test-retest reliability of $r = .88$ for repeated presentations of identical tokens, indicating that perceptual judgments were stable and systematic rather than random. Individual differences among listeners accounted for approximately 15% of variance in accuracy, with some listeners demonstrating consistently superior performance across all vowel categories.

8. Discussion

The findings demonstrate that perceptual stability is closely tied to acoustic distinctiveness. Mixed-effects modeling reveals that listener categorization is robust to speaker variability but sensitive to reductions in acoustic contrast. These results support exemplar-based models and challenge strictly categorical views of vowel perception (Johnson, 2006; Pierrehumbert, 2001).

The strong correlation between acoustic distance and perceptual accuracy provides compelling evidence for a direct production-perception linkage. Vowels that occupy peripheral positions in acoustic space, maximally distant from competing categories, are perceived with high accuracy and minimal confusion. Conversely, vowels in crowded regions of the vowel space, where multiple categories converge, show degraded perceptual discriminability. This pattern is consistent with the principle of acoustic-perceptual optimization, which holds that phonological systems evolve to maximize acoustic distinctiveness among contrasting categories.

The substantial inter-speaker variation documented in the acoustic analysis raises important questions about normalization mechanisms in speech perception. Despite considerable absolute differences in formant frequencies across speakers, listeners maintained high identification accuracy for most vowel categories. This suggests that perceptual normalization operates effectively to extract speaker-independent vowel quality information from variable acoustic input. The mixed-effects modeling approach employed in this study provides a framework for quantifying both systematic variation (captured by fixed effects) and random variation (captured by random effects), offering insights into the structure of variability that listeners must navigate.

The finding that female speakers produced expanded vowel spaces compared to male speakers, even after normalization, is consistent with several previous studies documenting gender-based differences in vowel production that exceed simple anatomical scaling. Several explanations have been proposed for this phenomenon, including differences in articulatory precision, learned phonetic targets associated with gender identity, and biomechanical constraints that vary between male and female vocal tracts. Further research employing articulatory methods such as ultrasound or electromagnetic articulography would be valuable for distinguishing among these competing hypotheses.

The role of vowel duration as a secondary cue for vowel identification highlights the multi-dimensional nature of vowel perception. While formant frequencies are unquestionably the primary acoustic correlates of vowel quality, temporal information contributes to perceptual robustness, particularly for vowel pairs with overlapping spectral characteristics. This finding aligns with models of cue integration in speech perception, which propose that listeners weight multiple acoustic dimensions probabilistically in making phonetic judgments.

From a theoretical perspective, the results provide strong support for exemplar-based models of speech perception. The gradient relationship between acoustic distance and perceptual accuracy, rather than categorical boundaries with sharp perceptual transitions, is more consistent with exemplar approaches than with traditional structuralist models positing discrete phonemic categories. Additionally, the substantial

individual variation among both speakers and listeners documented in the mixed-effects analyses aligns with the exemplar view that linguistic representations are shaped by personal experience and retain detailed phonetic information.

The implications of these findings extend to applied domains including speech technology and second language acquisition. Automatic speech recognition systems must contend with the same variability that human listeners navigate successfully, suggesting that incorporation of normalization mechanisms and probabilistic category representations could improve system performance. In second language learning, understanding the acoustic-perceptual relationships documented here can inform pedagogical approaches, highlighting the importance of establishing acoustically distinct phonetic categories to support perceptual learning.

9. Conclusion

By integrating acoustic analysis, perception experiments, and advanced statistical modeling, this study provides a robust account of vowel variation. The use of mixed-effects models strengthens the empirical validity of the findings and aligns the study with current best practices in experimental phonetics.

The research demonstrates that vowel production exhibits systematic variation across speakers while maintaining sufficient acoustic distinctiveness to support reliable perceptual categorization. Acoustic distance in F1-F2 space emerges as a key predictor of perceptual accuracy, indicating that the organization of vowel space directly impacts perceptual discriminability. These findings bridge the production-perception divide, showing that acoustic patterns arising from articulatory constraints have direct perceptual consequences.

The mixed-effects modeling framework employed in this study offers significant advantages over traditional analytical approaches by explicitly accounting for multiple sources of variation in speech data. This methodology enables more accurate estimation of experimental effects while also providing quantitative insights into the structure of variability in both production and perception. As computational tools for mixed-effects modeling become increasingly accessible, their adoption in experimental phonetics will continue to enhance the rigor and replicability of research findings.

Future research should extend these findings in several directions. First, longitudinal studies tracking vowel production and perception across development would illuminate how acoustic-perceptual relationships are established during language acquisition. Second, cross-linguistic comparisons examining vowel systems with different densities and organizations would test the generalizability of the production-perception mappings documented here. Third, investigations incorporating additional acoustic dimensions such as formant dynamics and spectral tilt would provide a more complete characterization of the acoustic information supporting vowel perception.

Additionally, research employing neuroimaging methods to investigate the neural correlates of vowel perception would complement the behavioral findings reported here, potentially revealing the neural mechanisms underlying perceptual normalization and category formation. Finally, computational modeling using neural network architectures trained on naturalistic speech input could test whether artificial systems develop perceptual strategies similar to those employed by human listeners, providing convergent evidence for the principles governing speech perception.

In conclusion, this study advances our understanding of vowel variation by demonstrating systematic relationships between acoustic production patterns and perceptual categorization accuracy. The integration of experimental phonetic methods with sophisticated statistical modeling provides a methodological template for future research in speech science, highlighting the value of quantitative, data-driven approaches to longstanding questions in phonetics and phonology.

References

1. Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects. Journal of Memory and Language, 59, 390–412.
2. Boersma, P., & Weenink, D. (2024). Praat: Doing phonetics by computer [Computer program].
3. Bybee, J. (2001). Phonology and language use. Cambridge University Press.
4. Fant, G. (1960). Acoustic theory of speech production. Mouton.
5. Harrington, J. (2010). Acoustic phonetics and sound change. Journal of Phonetics, 38, 203–214.
6. Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception. Journal of Phonetics, 34, 458–484.
7. Hillenbrand, J., Getty, L., Clark, M., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. Journal of the Acoustical Society of America, 97, 3099–3111.
8. Iverson, P., & Kuhl, P. (1995). Mapping the perceptual magnet effect. Journal of the Acoustical Society of America, 97, 553–562.
9. Johnson, K. (2006). Exemplar models in phonetics. Journal of Phonetics, 34, 485–499.
10. Johnson, K. (2012). Acoustic and auditory phonetics. Wiley-Blackwell.
11. Ladefoged, P. (2003). Phonetic data analysis. Blackwell.
12. Ladefoged, P., & Johnson, K. (2015). A course in phonetics. Cengage.
13. Nearey, T. (1989). Static, dynamic, and relational properties. Journal of the Acoustical Society of America, 85, 2088–2113.
14. Peterson, G., & Barney, H. (1952). Control methods used in vowel studies. Journal of the Acoustical Society of America, 24, 175–184.
15. Pierrehumbert, J. (2001). Exemplar dynamics. In Frequency and the emergence of linguistic structure.
16. Stevens, K. (1998). Acoustic phonetics. MIT Press.
17. Winter, B. (2019). Statistics for linguists. Routledge.
18. Keating, P. (1988). Underspecification in phonetics. Phonology, 5, 275–292.
19. Cho, T., & Ladefoged, P. (1999). Variation and universals in VOT. Journal of Phonetics, 27, 207–229.
20. Thomas, E. R. (2011). Sociophonetics. Palgrave.
21. Shosted, R. (2006). Vowel space typology. Journal of Phonetics, 34, 219–241.
22. Kent, R. D., & Read, C. (2002). The acoustic analysis of speech. Singular.
23. Fletcher, J. (2010). Prosody and speech timing. Laboratory Phonology, 1, 523–602.
24. Gordon, M. (2016). Phonological typology. Oxford University Press.
25. Wells, J. C. (1982). Accents of English. Cambridge University Press.