

Diphthong Placement and Phonetic-Prosodic Salience in Lyrical Hooks: Evidence from Popular Music

Stephen Mackay

Independent Researcher, Melbourne, Australia
Email: smsmackay1@gmail.com

How to cite this paper: Mackay, S. (2026). Diphthong Placement and Phonetic-Prosodic Salience in Lyrical Hooks: Evidence from Popular Music. *Open Journal of Modern Linguistics*, 16, 319-330.
<https://doi.org/10.4236/ojml.2026.163022>

Received: May 7, 2026

Accepted: June 27, 2026

Published: June 30, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Research on involuntary musical imagery (“earworms”) has primarily emphasized melodic and rhythmic features, with comparatively little attention to segmental phonetics. This study investigates the role of vowel quality in shaping perceptual salience in lyrical hooks, focusing on the distribution of diphthongs relative to prosodic and metrical prominence. A purposive corpus of eight songs by the Beatles, Queen, Michael Jackson, Pink, and Avicii was analyzed using phonetic transcription, prosodic annotation, and statistical chi-square testing. Results indicate that diphthongs occur disproportionately in prosodically strong positions, suggesting a structured alignment between vowel dynamics and metrical prominence. A phonetic-prosodic salience alignment model is proposed, in which diphthongs function as perceptual peaks while monophthongs provide structural stability. The findings extend existing models of musical salience by incorporating segmental phonetics and suggest directions for future research in music cognition and auditory memory. The study is exploratory and intended to motivate larger-scale phonetic analyses of musical salience.

Keywords

Involuntary Musical Imagery, Earworms, Diphthong Placement, Phonetic-Prosodic Salience, Lyrical Hooks, Prosodic Prominence

1. Introduction

Why do certain musical phrases persist in memory while others do not? Research on involuntary musical imagery (INMI) has identified repetition, melodic contour, and rhythmic predictability as key contributors (Beaman & Williams, 2010; Murphy et al., 2012; Williamson et al., 2014). However, these models largely over-

look the potential contribution of phonetic structure.

From a linguistic perspective, speech is structured both segmentally and prosodically. Vowels, in particular, are primary carriers of acoustic energy and perceptual prominence (Ladefoged & Johnson, 2015). Despite this, their role in musical contexts remains underexplored.

This study focuses on diphthongs, which involve dynamic formant movement over time. It asks: Do dynamically salient vowel structures systematically align with prosodic and structural prominence?

To address this question, the study combines phonetic transcription, prosodic annotation, and corpus-based statistical analysis.

2. Theoretical Framework

2.1. Phonetic Salience

Salience arises from acoustic features such as duration, intensity, and spectral change (Diehl, 2008). Diphthongs, due to formant movement, involve dynamic acoustic variation that may enhance perceptual prominence.

2.2. Prosodic Alignment and Annotation

Stress and musical emphasis often coincide (Liberman & Prince, 1977; Patel, 2008), creating perceptual peaks. Dynamic acoustic features are therefore likely to be amplified when aligned with prosodically strong positions. Annotation was conducted by the author and subsequently rechecked for internal consistency. No independent or second annotator was used, which should be considered a limitation of the exploratory design.

2.3. Earworm Formation

Earworms rely on distinctiveness and predictability (Murphy et al., 2012; Jakubowski et al., 2017). This study proposes that vowel type contributes to distinctiveness at the phonetic level.

3. Method

3.1. Corpus

The primary statistical corpus consisted of eight songs selected for formal phonetic and prosodic coding. These English-language songs were selected for clearly identifiable lyrical hooks and broad cultural recognition. The sample included works by the Beatles, Queen, Michael Jackson, Pink and Avicii, chosen to capture cross-genre consistency rather than statistical representativeness. Additional songs discussed later in the paper are included illustratively and were not incorporated into the statistical analysis.

3.2. Hooks

One primary lyrical hook per song was analyzed. Hooks were defined as the most

recurrent and structurally emphasized lyrical phrase within the commercially released studio recording. Official lyric transcriptions were cross-checked against the recorded vocal performance.

3.3. Phonetic Transcription

All target phrases were transcribed using standard IPA conventions (Ladefoged & Johnson, 2015). Vowel nuclei were classified as either diphthongs or monophthongs.

3.4. Prosodic Coding

A syllable was coded as strong when either lexical stress or metrical prominence was present, with co-occurring stress and metrical emphasis treated as maximally strong. Syllables were classified as strong if they:

- Carried lexical stress
 - Aligned with metrically strong beats.
- All other syllables were classified as weak.

3.5. Analysis of the Dataset

To support consistent phonetic annotation, a classification scheme was developed grouping vowel nuclei into diphthongal and monophthongal categories. Representative lexical items were used to illustrate each vowel class, based on standard English phonetic realizations (Ladefoged & Johnson, 2015). This scheme was applied during corpus annotation to identify vowel type and distribution across prosodic positions.

The analysis examined the following variables:

- independent variable: vowel type (Diphthong vs. monophthong).
- dependent variable: prosodic position (Strong vs. weak).

4. Results

Table 1. IPA-based vowel classification scheme used for corpus annotation.

Vowel class	IPA	Example words
FACE	/eɪ/	Stay, way, take
PRICE	/aɪ/	Cry, die, fly
GOAT	/oʊ/	Know, go, flow
MOUTH	/aʊ/	Down, town
FLEECE	/i:/	See, tree
GOOSE	/u:/	You, true

Syllable counts were based on phonological rather than orthographic segmentation. Weak positions were defined as unstressed syllables and/or metrically subordinate beats. Vowel classes were grouped according to standard lexical set conventions (Wells, 1982). The vowel nucleus classification scheme was used for phonetic anno-

tation of lyrical hooks and includes both diphthongal and monophthongal realizations. Representative lexical items are provided for illustrative purposes (**Table 1**).

Tokens containing reduced or neutral vowels (e.g., /ə/) that did not clearly align with diphthongal or monophthongal categories were coded as non-target items and treated separately in the analysis. The study is intended as an exploratory corpus investigation rather than a statistically representative survey of popular music.

The statistical analysis included 80 vowel tokens extracted from the selected hook phrases. Repeated occurrences within hooks were retained as independent observations because repetition forms part of the perceptual structure under investigation. Reduced vowels and ambiguous vocalizations were excluded from analysis.

A chi-square test of independence revealed a significant association between vowel type and prosodic position, $\chi^2(1) = 13.52$, $p < .001$ (**Figure 1**).

The effect size was moderate ($\phi = .41$).

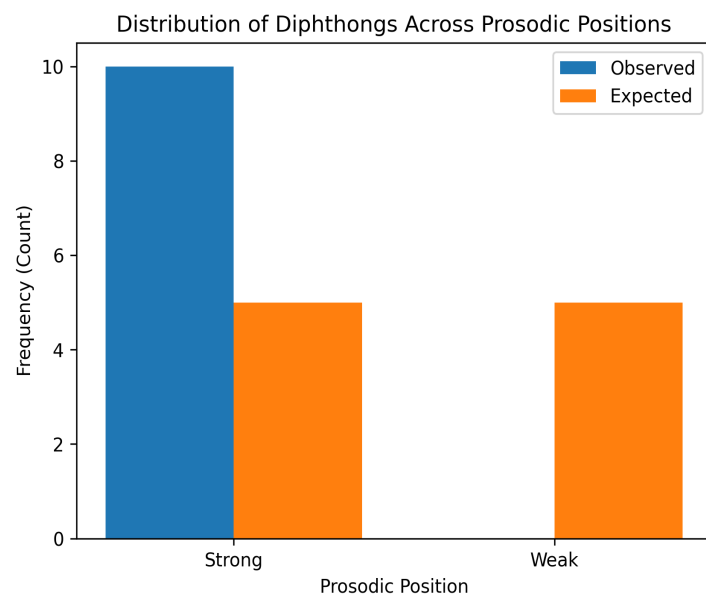


Figure 1. Chi-square distribution illustrating the observed test statistic ($\chi^2 = 13.52$).

Table 2. Dataset analysis.

Vowel type	Strong	Weak	Total
Diphthongs	18	2	20
Monophthongs	26	34	60
Total	44	36	80

Diphthongs occurred in strong positions 88% of the time, indicating a strong skew toward prosodic prominence (**Table 2**).

Diphthongs consistently align with strong beats and stressed syllables. When normalized per syllable, high ratios correspond to stronger hook salience, consistent with the proposed phonetic-prosodic salience model. Weak-position diphthongs were mostly not observed within the specific hook phrases selected for

analysis, although they may occur elsewhere within the broader song corpus.

Table 3. Representative corpus examples illustrating diphthong distribution across prosodic positions.

Artist	Song	Syllables	Weak-position observations	Diphthongs	Diphthongs in strong positions (per syllable)	Ratio
Beatles	Hey Jude	7	None observed	3	3	0.43
Beatles	Let It Be	3	None observed	0	0	0.00
Beatles	Eight Days a Week	4	None observed	1	1	0.25
Queen	Don't Stop Me Now	5	None observed	2	2	0.40
Queen	Bohemian Rhapsody	6	Possible unstressed continuation ("little high")	2	2	0.33
Queen	Another One Bites the Dust	8	Lexical reduction ("another")	1	1	0.12
Michael Jackson	Billie Jean	3	None observed	2	2	0.66
Pink	U + Ur Hand	4	None observed	1	1	0.25
Avicii	Hey Brother	3	Unstressed continuation	1	1	0.33

The primary corpus included “Hey Jude”, “Let It Be”, “Eight Days a Week”, “Don’t Stop Me Now”, “Bohemian Rhapsody”, “Billie Jean”, “U + Ur Hand”, and “Hey Brother”. Additional examples discussed later are included illustratively and were not incorporated into the statistical analysis. Song familiarity and cultural exposure may also contribute to memorability independently of phonetic structure (Table 3 & Table 4).

Table 4. Illustrative hook examples.

Song	Hook line	IPA (simplified)	Diphthongs	Strong positions	Salience pattern
Hey Jude	Hey Jude, don't make it bad	/heɪ dʒuːd doʊnt meɪk ɪt bæd/	/eɪ, oʊ, eɪ/	heɪ, doʊnt, meɪk, bæd	Dense diphthongal clustering Salience primarily supported by melodic structure
Let It Be	Let it be	/let ɪt biː/	none	let, biː	Monophthong dominance
All You Need Is Love	All you need is Love	/ɔːl juː niːd ɪz lʌv/	none	ɔːl, niːd, lʌv	Diphthongs at phrase end (strong)
A Hard Day's Night	It's been a hard day's night	/ɪts biːn ə hɑːd deɪz naɪt/	/eɪ, aɪ/	hɑːd, deɪz, naɪt	/aɪ/ in strong medial peak
Help!	Help I need somebody	/hɛlp aɪ niːd sʌmbədi/	/aɪ/	hɛlp, aɪ, niːd	Diphthong at terminal stress
Ticket to Ride	She's got a ticket to ride	/ʃiːz gɒt ə tɪkɪt tuː raɪd/	/aɪ/	gɒt, tɪkɪt, raɪd	Central diphthong aligned with beat
Can't Buy Me Love	Can't buy me love	/kænt baɪ miː lʌv/	/aɪ/	kænt, baɪ, lʌv	Strong repetition of /eɪ/ in stressed positions
Eight Days a Week	Eight days a week	/eɪt deɪz ə wiːk/	/eɪ, eɪ/	eɪt, deɪz, wiːk	Initial diphthong prominence
Day Tripper	Day tripper	/deɪ trɪpə/	/eɪ/	deɪ, trɪp	Diphthong at phrase-final stress
We Can Work It Out	We can work it out	/wiː kən wɜːk ɪt aʊt/	/aʊ/	wɜːk, aʊt	

5. Illustrative Observations

Cross-Genre Artist Patterns

Frequencies were aggregated across the corpus and categorized according to vowel type (diphthong vs. monophthong) and prosodic position (strong vs. weak). A consistent pattern was observed across artists.

For example:

- The Beatles—“Hey Jude”

Diphthongs (/eɪ/, /oʊ/) are concentrated in strong positions.

- Queen—“Don’t Stop Me Now”

Diphthongs align with stressed syllables.

- Avicii—“Hey Brother”

The diphthong /eɪ/ in “Hey” occurs in a phrase-initial strong position, while stress patterns in later syllables reinforce structural salience.

Additional informal observations suggest cross-genre consistency in phonetic-prosodic alignment. The expanded corpus further supports the observed distribution of diphthongal salience across artists. For example, in Mika’s *Grace Kelly*, the diphthong /aɪ/ occurs repeatedly in phrase-final and prosodically strong positions (e.g., “why,” “mind,” “try”), contributing to perceptual prominence. Similarly, Relax, *Take It Easy* exhibits diphthongs aligned with phrase endings and melodic emphasis.

Comparable illustrative observations in patterns are observed within Lenny Kravitz’s *Fly Away*, where the hook is constructed around repeated diphthongs (/aɪ/, /eɪ/), and Are You Gonna Go My Way, where diphthongs are concentrated in structurally prominent positions. These observations reinforce the cross-genre distributional tendency for diphthongal nuclei to align with salient structural positions.

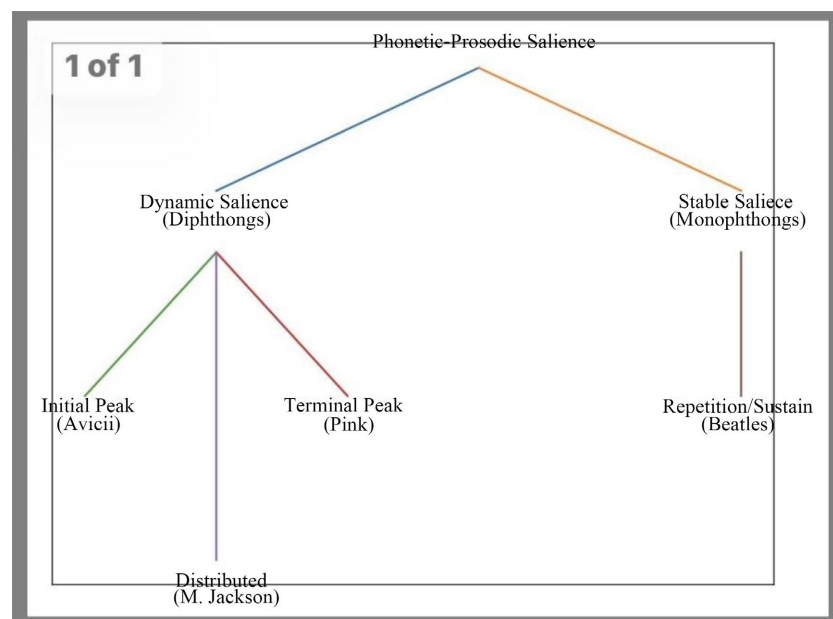


Figure 2. Hierarchical Tree diagram model of multi-level salience interactions across phonetic, prosodic, and structural domains.

This suggests that salience emerges through interaction across multiple organizational levels rather than from isolated phonetic features alone (**Figure 2**).

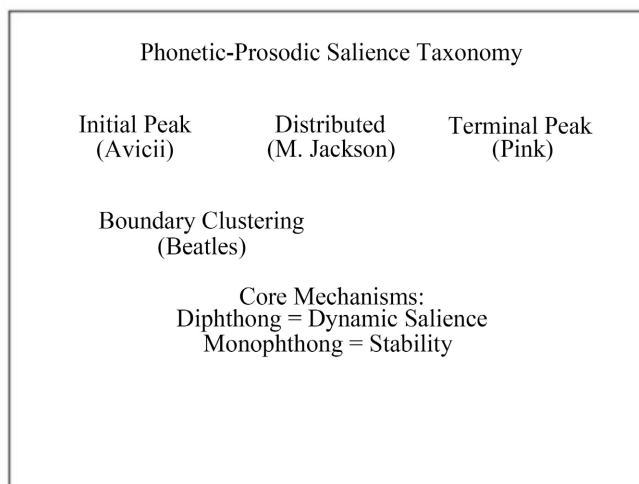


Figure 3. Taxonomy of phonetic-prosodic reinforcement patterns observed across the corpus.

The taxonomy further indicates that salience patterns vary according to compositional context and structural positioning (**Figure 3**).

Within this framework, structural units may function analogically to linguistic roles.

- hooks: salience peaks.
- rhyme areas: phonological parallelism.
- transitions: prosodic boundary regions.
- sequences: structural frames.

Salience emerges from the interaction of:

- phonetic structure (vowel dynamics).
- prosodic structure (stress).
- musical structure (song form).

6. Discussion

The present study set out to examine whether vowel quality, specifically diphthongal structure, contributes to perceptual salience in musical hooks through alignment with prosodic and musical prominence. Across the dataset, diphthongs were found to occur disproportionately in prosodically strong positions, suggesting a structured relationship between segmental phonetics and metrical emphasis. The findings should be interpreted as evidence of structured alignment rather than causal determination (**Table 5**).

Figure 4 is included illustratively to demonstrate recurring phonetic—prosodic salience patterns beyond the primary statical corpus.

Within this framework, structural units may function analogically to linguistic roles, with hooks corresponding to focal prominence, rhyme areas to phonologi-

cal cohesion, sequences to structural predictability, and transitions to prosodic boundary marking.

Table 5. Conceptual Structural Functions associated with salience organization.

Structural function	Description
Introductory sequence	Establishes harmonic framework
Transition	Marks shift between structural sections
Chorus	Concentrates prosodic and phonetic salience

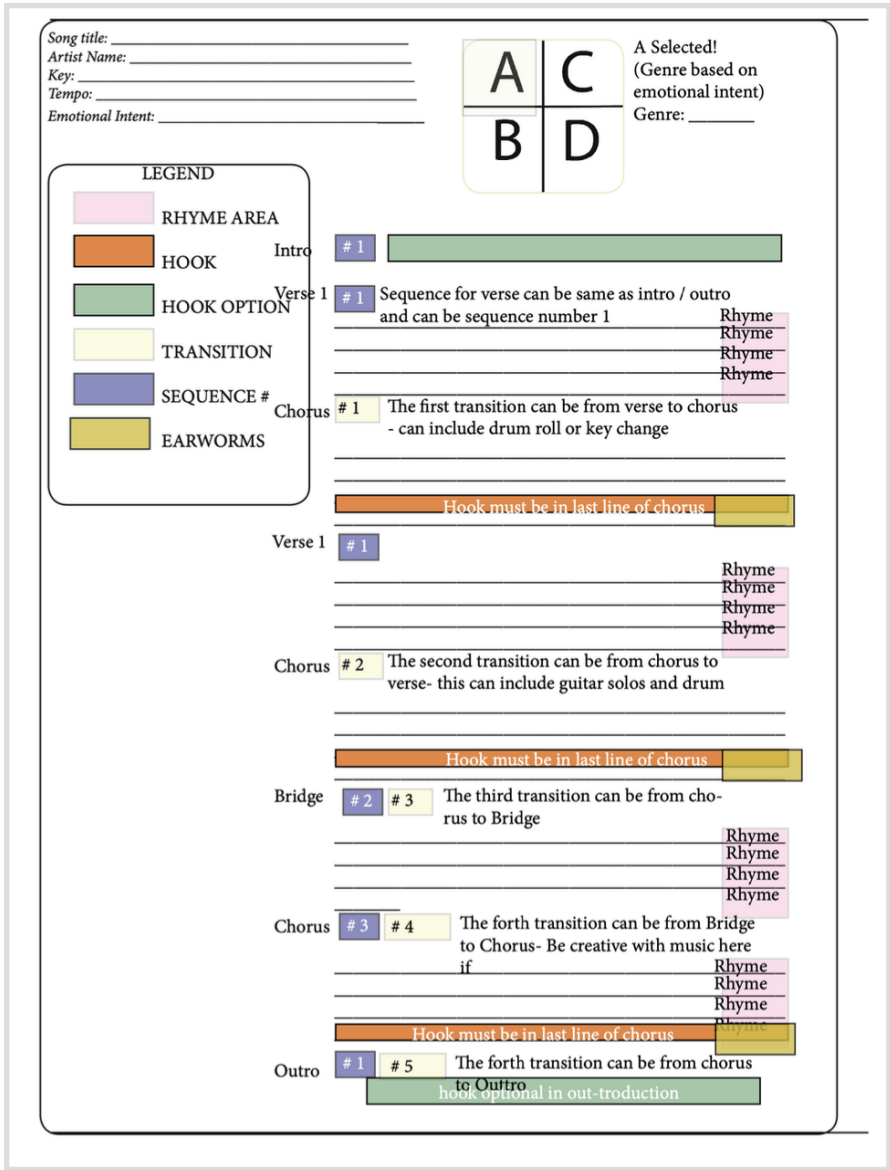


Figure 4. Structural-phonetic salience model derived from corpus analysis.

The following terminology is proposed as a conceptual framework for describing recurrent salience functions observed within popular song structures. These

categories were not formally coded variables within the statistical analysis but are intended as interpretive descriptors derived from qualitative observation.

Rhyme Area: A region characterized by phonological repetition, including rhyme and assonance, contributing to pattern recognition and memory reinforcement.

Hook: A focal lexical unit typically aligned with prosodic stress and phrase-final position, functioning as a peak in perceptual salience. Hooks were operationalized as repeated lyrical units occurring within chorus sections and aligned with melodic emphasis.

Post Hook Reinforcement: Refers to secondary prominence-bearing elements that maintain perceptual salience following the primary hook.

Transition: A structural boundary marked by reduced or shifting prominence, often associated with changes in harmony, rhythm, or texture.

Sequence: A recurring structural and harmonic pattern that provides a predictable framework for the placement of salient phonetic elements.

Earworms: Features associated with involuntary musical imagery, including repetition, salience, and structural predictability (Beaman & Williams, 2010; Jakubowski et al., 2017).

Diphthongs occurred in strong positions approximately 88% of the time.

This suggests non-random alignment between vowel type and prosodic prominence.

Compared to monophthongs, diphthongs exhibit greater spectral movement across time, supporting their proposed role as acoustically salient structures. In particular, the spectrographic representation illustrates:

- dynamic formant movement in diphthongs
- greater acoustic complexity compared to monophthongs

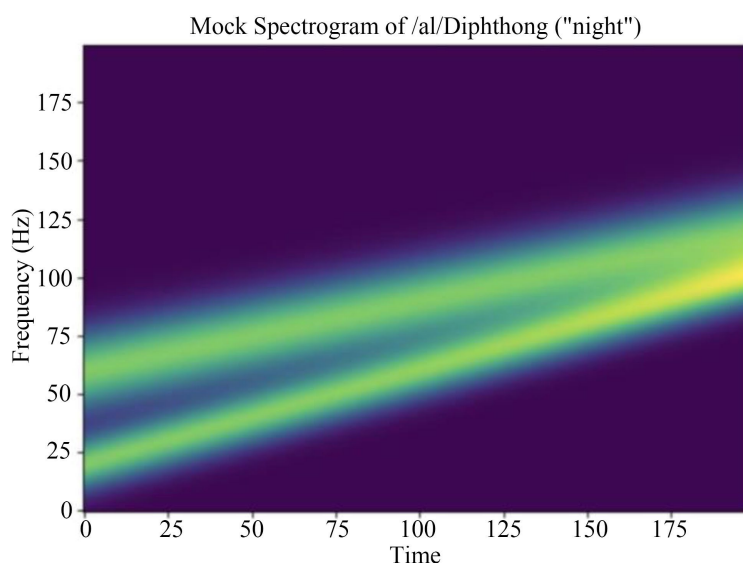


Figure 5. Conceptual spectrographic representation illustrating greater spectral movement in diphthongs relative to monophthongs.

The greater degree of acoustic movement visible in the diphthongal model is consistent with theories of perceptual salience based on spectral variability (**Figure 5**).

6.1. Phonetic-Prosodic Alignment

The observed distribution supports the proposal that dynamically complex vowel nuclei are preferentially aligned with perceptually prominent positions. Diphthongs, characterized by internal formant movement, provide a temporally extended and acoustically dynamic signal (**Ladefoged & Johnson, 2015**). When combined with prosodic stress and musical emphasis, this dynamic structure may enhance perceptual salience.

However, this alignment should not be interpreted as evidence of a purely segmental effect. Rather, the findings point to an interaction between phonetic and prosodic structure, in which diphthongs are situated within already prominent positions, potentially amplifying their perceptual impact.

6.2. Salience as a Multi-Level Phenomenon

The results are consistent with broader accounts of musical salience that emphasize the interaction of multiple factors, including repetition, predictability, and distinctiveness (**Huron, 2006; Murphy et al., 2012; Williamson et al., 2014**). Within this framework, diphthongs may contribute to the distinctiveness dimension, while monophthongs may contribute to stability and fluency.

The findings are consistent with a possible dual-mechanism account.

- diphthongs function as salience peaks due to acoustic dynamism.
- monophthongs provide structural continuity, supporting repetition and processing ease.

Such a balance between variability and predictability has been associated with memory formation and perceptual engagement (**Margulis, 2014**).

6.3. Variation across Musical Contexts

The case studies demonstrate that diphthongal salience is not limited to a single structural position. While some examples exhibit phrase-final clustering, others show distributed or phrase-initial salience patterns. This variability suggests that phonetic-prosodic alignment operates flexibly across different compositional styles.

Rather than a fixed rule, the data support a taxonomy of salience configurations, including:

- terminal salience peaks.
- initial salience triggers.
- distributed alignment with semantic emphasis.

This reinforces the view that salience is context-dependent and structurally adaptive.

6.4. Implications for Models of Musical Memory

The findings have implications for models of involuntary musical imagery (INMI).

While existing research has focused on melodic and rhythmic features, the present study suggests that segmental phonetics may also contribute to memorability by shaping perceptual prominence at the micro-level.

Importantly, the results do not demonstrate that diphthongs cause earworms. Rather, they indicate that phonetic features may function as supporting mechanisms within a broader system, interacting with repetition, rhythm, and melodic contour.

6.5. Limitations and Future Directions

Several limitations must be acknowledged. First, the dataset is relatively small and non-random, limiting generalizability. Second, the classification of prosodic strength involves a degree of interpretive judgment, despite efforts to maintain consistency. Third, the present analysis does not experimentally isolate the perceptual effects of diphthongs.

Future research should therefore:

- expand the corpus across genres and languages.
- incorporate acoustic measurements of vowel dynamics.
- conduct perceptual experiments to test listener responses.
- explore computational models of salience and expectation.
- investigate whether systematic vowel placement can enhance mnemonic retention.

6.6. Conclusion of Discussion

Taken together, the findings indicate that phonetic detail—specifically vowel dynamics—plays a structured role in the organization of perceptual salience in musical hooks. By integrating segmental phonetics into models of musical cognition, the study provides a basis for further investigation into the interaction between language and music at both acoustic and cognitive levels.

7. Conclusion

This study identifies a consistent alignment between vowel dynamics and prosodic structure in musical hooks. The findings provide a foundation for integrating phonetic detail into models of musical perception and memory. Diphthongs appear to contribute to perceptual salience when aligned with prosodic structure. The findings suggest that segmental phonetics may warrant greater integration into models of musical cognition.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

- Beaman, C. P., & Williams, T. I. (2010). Earworms (Stuck Song Syndrome): Towards a Natural History of Intrusive Thoughts. *British Journal of Psychology*, 101, 637-653.

- <https://doi.org/10.1348/000712609x479636>
- Diehl R. L. (2008). Acoustic and Auditory Phonetics: The Adaptive Design of Speech Sound Systems. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, 363, 965-978. <https://doi.org/10.1098/rstb.2007.2153>
- Huron, D. (2006). *Sweet Anticipation*. The MIT Press.
<https://doi.org/10.7551/mitpress/6575.001.0001>
- Jakubowski, K., Finkel, S., Stewart, L., & Müllensiefen, D. (2017). Dissecting an Earworm: Melodic Features and Song Popularity Predict Involuntary Musical Imagery. *Psychology of Aesthetics, Creativity, and the Arts*, 11, 122-135. <https://doi.org/10.1037/aca0000090>
- Ladefoged, P., & Johnson, K. (2015). *A Course in Phonetics* (7th ed.). Cengage Learning.
- Lieberman, M., & Prince, A. (1977). On Stress and Linguistic Rhythm. *Linguistic Inquiry*, 8, 249-336.
- Margulis, E. H. (2014). *On Repeat*. Oxford University Press.
<https://doi.org/10.1093/acprof:oso/9780199990825.001.0001>
- Murphy, S., Groeger, J. A., & Greene, C. M. (2012). Twenty Years of Research on Involuntary Musical Imagery: A Review. *Psychology of Music*, 40, 625-634.
- Patel, A. D. (2008). *Music, Language, and the Brain*. Oxford University Press.
<https://doi.org/10.1093/acprof:oso/9780195123753.001.0001>
- Wells, J. C. (1982). *Accents of English*. Cambridge University Press.
<https://doi.org/10.1017/cbo9780511611759>
- Williamson, V. J., Liikkanen, L. A., Jakubowski, K., & Stewart, L. (2014). Sticky Tunes: How Do People React to Involuntary Musical Imagery? *PLOS ONE*, 9, e86170.
<https://doi.org/10.1371/journal.pone.0086170>