5-2022

In Search of Phonetic Evidence for Prosodically-Motivated Aspiration

McKinley Sprinkle
William & Mary

In Search of Phonetic Evidence of Prosodically-Motivated Aspiration

A thesis submitted in partial fulfillment of the requirement
for the degree of Bachelor of Arts in Linguistics from
William & Mary

by

McKinley T. Sprinkle

Accepted for Honors

Dr. Anya Hogoboom, Director

Dr. Daniel Parker

Dr. Kaitlyn Harrigan

Dr. Erin Webster

Williamsburg, VA
May 1, 2022

Acknowledgements

As with any research project of this scope, this thesis would not have been possible without the assistance of numerous people at various stages of the process.

I am especially indebted to Anya Hogoboom, my thesis supervisor, for her consistent help on this project since October of 2020. Her knowledge of linguistic literature and experimental best practice, regular brainstorming, and consistent feedback across a year and a half has been invaluable, and this project would not have been possible without her diligence, enthusiasm, and encouragement.

I am also extremely grateful to Kate Harrigan and Kim Love, whose feedback and extensive input on experimental design and statistical analysis were crucial to the execution of several of the studies in this paper.

The William & Mary Linguistics research group, headed by Dan Parker, was also a wonderful source of feedback for this project, challenging me with difficult questions and giving me an opportunity to present various forms of this research on multiple occasions. Dr. Parker's contributions in particular helped to expand the analytical approaches used on this project and were extremely helpful.

I would also like to thank Erin Webster, whose gracious participation in this project as an outside member of my thesis committee and feedback as a non-expert audience was incredibly helpful in encouraging me to write for a more general audience, rendering a highly technical thesis at least somewhat decipherable.

I am also very grateful to the 159 people who participated in my studies in various forms, who generously provided their conscious and unconscious expertise in the English language.

Lastly, I must express my gratitude to the four anonymous reviewers for WCCFL 40, whose comments on the ongoing project helped tie it to existing literature and direct attention to shortcomings in the experimental process.

In Search of Phonetic Evidence for Prosodically-Motivated Aspiration

McKinley T. Sprinkle

Abstract: This thesis examines the production and perception of aspiration in all possible levels of stress and word positions attested under the left-edge prosodic description theorized by Kiparsky (1979), Withgott (1982), and Jensen (2000), as well as in all attested environments for unaspirated voiceless stops. Through the metric of voice onset time (VOT), I phonetically test the realization of aspiration and examine its perception as categorical in several environments that are not acoustically salient. Through a production study and two linked perception studies I provide acoustic evidence in support of the phonological definition of categorical aspiration as prosodically-motivated in English, and clarify the behavior of aspiration in two related stress lapse environments.

1.0 Introduction

As an acoustic phenomenon with predictable yet complex realizations, aspiration is directly in the intersection between the linguistic disciplines of phonetics and phonology. Phonetics is concerned with the physical properties of sound, their manifestations in the speech stream, and the physiological means of pronunciation. Phonology is concerned with the mental encoding of language, describing natural language through theoretical rules and features, and seeking to codify the way that humans innately use and construct language.

Aspiration is a phonological feature present in many languages as a way to distinguish different categories of stop; in English, it helps differentiate between voiced and voiceless stops, such as [b] and [p]. As a phonological feature, it is subject to and described by the grammar of a language, codified in phonological rules that accurately and consistently predict the behavior of that feature. While aspiration in English has historically been difficult to tie to only one unified phonological rule, Kiparsky (1979) and Withgott (1982) unified the multiple rules for where it can surface, using prosodic structure to concisely describe all of the environments in which it appears. Their theoretical analyses rely on conventionally-described environments for aspiration rather than their phonological descriptions being based on production data from speakers.

In contrast to the robust and elegant prosodic description of aspiration posited in the phonological literature, phonetic studies on aspiration have been quite limited in scope, principally focusing on short words and word-initial positions (e.g. Jiang et al. 2006, Kim et al. 2018). To my knowledge, the breadth of environments for aspiration described under the prosodically-motivated rule have not been considered in production studies, nor has the realization of aspiration in all of the possible conditions under the united prosodic rule been empirically tested.

This study seeks to provide phonetic analyses of aspiration in support of the existing phonological theory through the use of a production task and two perception tasks, one targeting categorical and the other gradient perception of aspiration, and to clarify lingering ambiguities in the application of the unified rule for aspiration in English.

2.0 Background

2.1 Phonetic understanding of aspiration

Aspiration is a burst of air that accompanies a stop consonant in some languages, and in English is present with the pronunciation of voiceless stops in many environments. Aspiration is one of the phonetic features that helps distinguish the sounds [p], [t], and [k] from their voiced equivalents [b], [d], and [g]. The prominence of aspiration present in the pronunciation of a consonant is found by measuring voice onset time (VOT), the time difference between the release of the burst of air and the beginning of the voicing of the vowel following the consonant. As shown in Figure 1, voiceless aspirated stops such as [p] have a notable positive VOT, while voiced unaspirated stops like [b] do not. Aspiration is such an important cue to the identity of voiceless stops that artificially reducing the amount of VOT present before a recorded voiceless stop to a near-zero level can result in the sound being perceived as a voiced stop without any other acoustic changes from the voiceless original pronunciation.

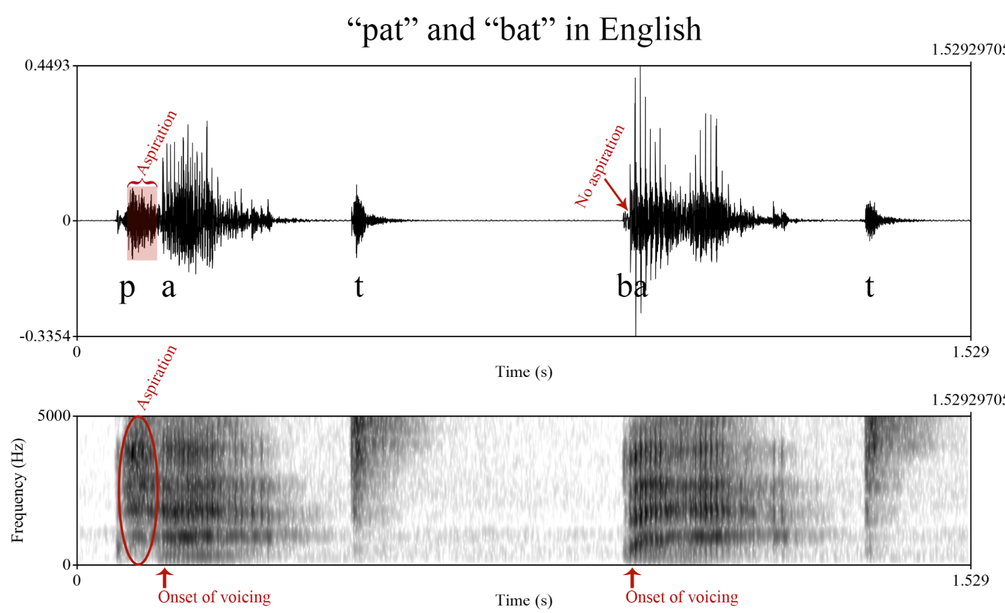

Figure 1: Waveform and spectrogram of "pat" and "bat," showing the presence of VOT prior to the [p] in "pat" and its absence prior to the [b] in "bat"

Measurement of VOT is not the only method of defining and examining aspiration in the field of linguistics; additional factors include F0 height, as in Whalen et al. 1993, or amplitude of aspiration, as in Repp 1979, both of which cause phonetically shorter VOT durations to be perceived as more significantly aspirated. VOT is also a somewhat limited metric for cross-linguistic use because of the existence of languages that use voicing and aspiration as independent features, such as Hindi, which has a four-way voicing and aspiration distinction (Abramson and Whalen 2017). Yet because English has only a two-way distinction between aspirated and unaspirated and the effect of other features in its perception can be accounted for, VOT functions as a reliable and easily measurable way to evaluate the presence of aspiration for voiceless stops and to distinguish them from voiced stops (Abramson and Whalen 2017).

2.2 Phonological understanding of aspiration

Even though it is phonetically realized as voice onset time, with technically infinite variation in its magnitude, aspiration is known to be categorical. In English, it is phonologically encoded in the mental grammar as either on or off, where large amounts of VOT correspond to aspiration and small or negative amounts of VOT (prevoicing) correspond to non-aspiration. Abramson and Whalen (2017) note that the presence of a small amount of VOT prior to a voiceless stop in English is not sufficient to perceptually mark it as aspirated, but that a more substantial amount of VOT is needed to be phonologically perceived as aspiration. This denotes that there is some threshold at which the amount of VOT becomes enough for listeners to perceive a sound as aspirated, below which it is perceived as not being aspirated.

Languages utilize different VOT values for the threshold between unaspirated and aspirated stops (Cho and Ladefoged 1999). There is no clear consensus on an exact VOT length of the aspiration threshold in English; Reetz and Jongman (2009) and Johnson (2012) both place it somewhere in the realm of 30ms-35m. Even in the absence of a precise measurement, which could be expected to vary based on regional variation in the same way that it does between languages (as described by Cho and Ladefoged 1999 and Abramson and Whalen 2017), this is generally recognized to be the range of duration in which voiceless stops with positive VOT perceptually change from unaspirated to aspirated in English.

2.3 Rule-based vs prosodic analyses of aspiration

Aspiration in English has posed a challenge to linguistic theory because it occurs in a large number of environments, seeming to defy the expectation that environments for phonological processes are natural classes; that is, they are identified under a single specific description that applies in every case across the language. Davis and Cho (2003) list the following environments in which aspiration can occur:

(1)

a. At the beginning of a syllable with primary stress (*conduit, repeater, tapestry*)
b. At the beginning of a syllable with secondary stress (*kangaroo, paradoxical, advertise*)
c. At the beginning of a word-initial stressless syllable (*cabana, potato, tacoma*)
d. At the beginning of a stressless syllable when immediately preceded by a stressless syllable and followed by a stressed one (*abracadabra, lollapalooza, Mediterranean*)

Kiparsy (1979) asserts that aspiration may be described as occurring at the left edge of a metrical foot. In prosody, feet are a rhythmic structure that relate to stress patterns. A foot is composed of a stressed syllable along with some number of stressless adjacent syllables (Hammond 2020). English employs trochaic feet, consisting of a strong syllable followed by a weak syllable, which generally results in stressed syllables occurring on the left edge of feet, as Kiparsky posits. This description of aspiration accounts for environments (1a) and (1b) above. The examples below show the correlation of aspiration and footing in both of these environments. Footing is shown using parentheses, with periods denoting syllable breaks and CV (or CVC) representing consonants and vowels within a syllable.

(2)

a. (CV́.CV)CV   b. (CV̀.CV)(CV́.CV)

Withgott (1982) added to the previously documented cases an additional medial environment for aspiration, (1d) above, which is explained by Withgott as well as Jensen (2000) as being a product of recursive foot structure, where a foot can be dominated by another “superfoot” for the purposes of prosody, thereby adjoining an additional syllable to either the left or the right of the original foot. This allows syllables that are structurally the left adjunct of a foot to also be foot-initial, therefore falling under the description of aspiration as occurring at the left edge of some level of foot, though Hammond (2020) argues that this is a controversial argument for recursing footing. If accurate, this description accounts for the environments listed above in (1c) and (1d). The examples below show the correlation of aspiration and superfoot boundaries for both of these environments.

(3)

a. (CV(CV́.CV)) b. (CV̀)(CV(CV́.CV))

The use of the left-edge definition of aspiration unifies all of these disparate environments into one natural class based on prosodic structure, aligning with the expectation that phonological processes are based on natural classes rather than multiple simultaneously-applying rules. Figure 2 below shows the recursive footing of each of these four environments for aspiration.

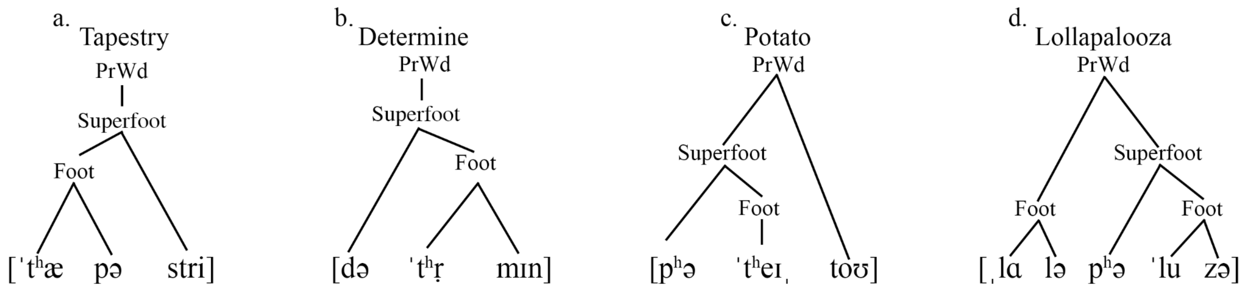

Figure 2: Recursive foot representations of the four rule-based environments for aspiration

With a clearer understanding of where aspiration is expected to occur, it is useful to understand the environments where voiceless stops may surface in English without being aspirated; that is, the environments where aspiration does not apply to voiceless stops yet in which they can still appear. Davis and Cho (2003) list the following as environments in which aspiration does not surface:

(4)

a. In coda position (*actress*, *lapse*, *atlas*)
b. At the beginning of a stressless syllable following a stressed one (*matriculation*, *rapid*, *nautical*)
c. As the possible second member of an onset consonant cluster (*skeleton*, *inspiration, statistics*)

In all of these environments, the voiceless stops do not appear at the left edge of a foot, aligning with the prosody-based description of aspiration. The examples below show the footing of a word in each of the environments in (4) above. The target sound that does not aspirate is underlined.

(5)

a. (CV́C.CV) b. (CV́.CV) c. (CCV́.CV)CV

One additional environment for unaspirated voiceless stops in English that has not to my knowledge been examined is that of final stress lapse, as in the word *harmonica.* This environment would not fall under any of the rule-based descriptions of aspiration or the left-edge description for multiple reasons: it is stressless, it is the final syllable in a foot, and it is at the right edge of a word, making it impossible to be the left edge of a foot, as shown in example (6) below:

(6)

a. (CV(CV́))CV.CV

2.4 Confounds to the foot-based theory of aspiration

Withgott (1982)'s assertion that the environment of voiceless stops in (1d) is aspirated is somewhat surprising given the disjunction in acoustic strength between aspiration and that environment. Aspiration is a phonologically strong feature, and in all of the other posited environments appears either on a stressed syllable or at the left edge of a word, both of which are perceptually salient positions in a word. By contrast, the environment of the [p] in words like *lollapalooza* (1d) is quite weak; it is not word-initial, it is not stressed, and it is in a stress lapse position, meaning that it is the second consecutive stressless syllable. The contrast between this very perceptually weak position and the proposed presence of a phonologically strong feature like aspiration is surprising.

While some study has been devoted to environments for voiceless stops in which aspiration does *not* occur, the lack of study of the final stress lapse condition identified in (6) above seems to be an oversight. While this environment would by no means be expected to aspirate give its right-edge position, its status of having a voiceless stop in the onset of a stress lapse syllable means that it is a somewhat close parallel to the environment posited as hosting aspiration in (1d), and the level of aspiration that surfaces in the target position seems perceptually similar to that of (1d), suggesting that an acoustic comparison between these two environments might be fruitful. Figure 3 below shows a comparison between the footing of these two environments, showing how differently they are treated prosodically.

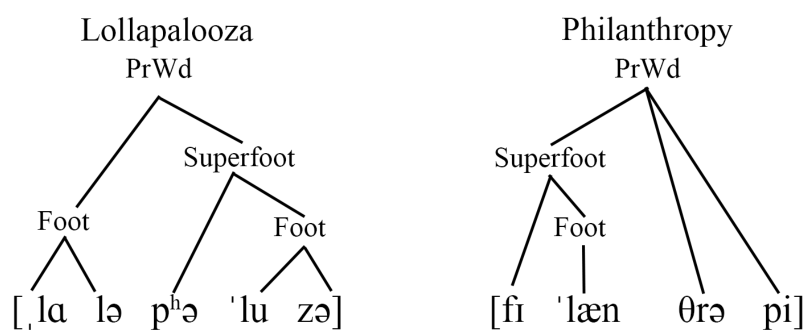

Figure 3: Comparison between word-medial and word-final stress lapse environment foot structures

Furthermore, McCarthy (1982) discusses the use of expletive infixation as a method of highlighting word-internal foot boundaries, and uses the environment described in (1d) as an example that has some flexibility in the location of expletive infixation. I show below how the word *lollapalooza* has two possible infixation environments: *lolla-fuckin-palooza* and *lollapa-fuckin-looza,* based on McCarthy's analysis of this stress environment. The first possibility utilizes the footing shown in Figure 2 above, where the antepenultimate syllable 'pa' in the root word is a leftward foot adjunct to the rightmost foot 'looza', thereby falling under the description for aspiration. However, an alternative footing of the word with the antepenult 'pa' acting as a rightward foot adjunct to the initial foot 'lolla' is also a possibility, and McCarthy states that this is slightly more acceptable to speakers than the alternative infixation location (1982:581). Both of these infixation locations are shown in comparison in Figure 4 below; the left footing structure aligns with Withgott (1982)'s description of aspiration on the antepenultimate syllable, and the right does not.

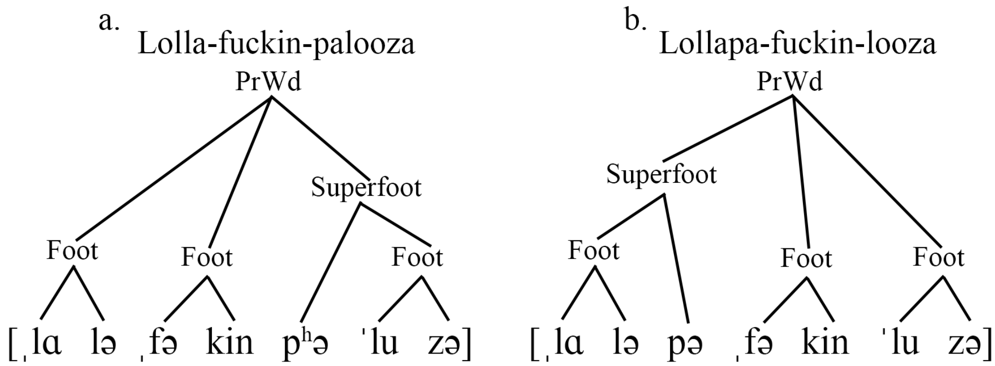

Figure 4: Footing diagrams of two possible expletive infixation locations in a medial stress lapse environment

Jensen (2000) rejects this ambiguous footing of the second word-medial stressless syllable, and motivates the rightward foot adjunction of the second word-internal stressless syllable by specifying that foot adjunction occurs first to the right and then to the left, resulting in the syllable ‘pa’ becoming a left foot adjunct to the final foot ‘looza,’ yet also permitting right foot adjuncts in specific environments. However this does not seem to account for the perceptual acceptability of *lollapa-fuckin-looza* that McCarthy (1982) discusses, which would imply the ‘pa’ syllable is a right foot adjunct to the initial foot, ‘lolla.’ If the medial stress lapse syllable is indeed a leftward foot adjunct to the final foot, then its initial stop should surface with aspiration. If, however, the stop in the onset of the stress lapse syllable does not carry aspiration, then its footing remains ambiguous, in defiance of Jensen’s description.

2.4 The current study

The purpose of the current series of studies is twofold, focusing on both examining the less strongly-attested phonological environments of aspiration and recording and comparing the phonetic realization of aspiration in as many different environments as possible.

On the first front, this study is concerned with examining the realization of aspiration in the medial stress lapse position, and determining whether the acoustically strong feature of aspiration does indeed manifest in the acoustically weak stress lapse position posited by Withgott (1982).

On the second front, this study is concerned with examining the phonetic levels of VOT that appear in all different aspiration environments. Outside of phonological and prosodic theory, little phonetic work has been done on acoustically measuring VOT in different word positions, especially beyond relatively short words (e.g. Kim et al. 2018, which focused on two-syllable words). The presence of aspiration in the third syllable of a five-syllable word, as posited in (1d) above, is well beyond the environments that have been phonetically examined, especially in comparison to the final lapse environment. This study will record the phonetic presence of aspiration in all environments and compare them to each other.

To accomplish these aims, the study consists of three experiments: one production study and two linked perception studies. The production study aims to record the amount of voice onset time present in every possible aspiration position, as well as every attested environment of voiceless stops in English that do not aspirate, to compare their acoustic manifestations. The two perception studies use recordings produced by participants in the production study as stimuli for a categorical and a gradient listening study, probing the manifestation of a specific aspiration threshold and gauging the phonetic sensitivity of English speakers to changes in VOT at different word positions.

Through the interaction of these three studies I hope to obtain and examine empirical evidence for the foot-based description of aspiration, determining whether the existing description adequately describes all known aspiration environments and excludes non-occurrence environments, or whether further refinement is needed in order to definitively tie aspiration to a prosodic description.

3.0 Experiment 1: Production

The production experiment is intended to measure the levels of voice onset time present for voiceless stops [p] and [k] in the recorded speech of participants at all possible stress levels and word positions, including positions where aspiration is not expected to surface. Through these measurements, I compare the relative strengths of various factors in influencing the presence and strength of aspiration, and create recordings that are used as stimuli for later perception studies.

While the production study also measures the VOT levels of [t] in these positions, /t/ is expected to surface as allophonic variants (glottal stop and flap) in certain environments, preventing the full analysis of [t] in relation to [p] and [k]. While it is included as a comparison case, this inability to analyze all possible environments means that it is not as rigorously examined as [p] or [k], test words with [t] in target positions are not as frequent in the word list used for this study (see Appendix A).

<u>3.2 Participants</u>

There were 25 participants for the production study. Participants were undergraduate students or recent alumni, and were compensated either with course research credit or monetarily with $5 at the conclusion of the study. The purpose of the study was explained to participants, and they were given the opportunity to withdraw at any time. Each participant read 23 question and answer pairs.

Of the 25 participants, 14 identified themselves as female, 10 as male, and 1 as non-binary. The mean age of participants was 19, with an age range of 18-25 years.

<u>3.1 Stimuli</u>

Words were selected for the production study to fit into 14 different possible environments; seven where aspiration was hypothesized to occur, and seven where it was not expected to. These environments were based on the aspiration conditions described by Davis and Cho (2003), but were expanded to incorporate variation based on stress level and word position and to include the environment of word-final lapse not considered in that or, to my knowledge, in previous descriptions of aspiration environments.

Two constraints directed word selection for all of these conditions: avoiding compound, affixed, or plural words in order to avoid significant morphological distinctions between syllables, and balancing the two main target consonants [p] and [k], with some inclusion of the comparison consonant [t], within each condition and across all conditions. Further, every attempt was made to select three-syllable words, or five-syllable words in specific conditions, in order to keep word lengths similar and matched to the most experimentally important study conditions. This syllable-length constraint was less strongly prioritized in conditions where aspiration was not present, as words matching the length criteria sometimes were not readily available in the language.

For each of the identified aspiration and nonaspiration conditions, a list of ten words was created, except in cases where five words were used, as noted and explained below. In conditions with ten words, the selected words were balanced to include 3 [t] sounds, and either 3 [p] and 4 [k] sounds or vice versa. In conditions with only five words, only one [t] sound was included, with 2 [p] and [k] sounds. As discussed more extensively in §3.0 above, [t] sounds were deprioritized throughout the word list as they are expected to become flaps or glottal stops in certain environments where aspiration is not expected.

There were a total of 115 test words in the word list. 42 words had [k] in the target position, [41] had [p], and [32] had [p]. A summary of each of the aspirations environments is provided in Table 1, and a summary of voiceless stop relizations without aspiration are provided in Table 2. A full listing of all test words is provided in Appendix A.

<u>3.1.1: Expected environments for aspiration</u>

The first aspiration environment Davis and Cho (see §2.2 above) identify is where the target sounds ([p], [t], and [k]) appear at the beginning of a syllable with primary stress, such as in *[k]onduit.* For this study, this environment has been broken into two separate conditions; Condition 1, where the primary stress occurs word-initially, as in *[k]aravan* and *[p]aradise,*' and Condition 2, where the primary stress occurs non-initially, as in *a[t]onic* and *re[p]eater.*'

The second aspiration environment is where target sounds occur at the beginning of a syllable with secondary stress, such as in *[t]angerine.* Like the first environment, this environment has been split into two conditions based on word position; Condition 3 consists of words where the secondary stress occurs word-initially, as in *[k]angaroo* and *[p]reconceive,* and Condition 4 consists of words where secondary stress occurs non-initially, as in *indi[k]ate* and *honey[k]omb*. Additionally, Condition 5 consists of a list of words where the secondary stress appears word-initially and is composed only of five-syllable words, to match with the fourth aspiration environment (Condition 7; see below), as in *[k]afeteria* and *[p]aradoxical.*

The third aspiration environment is where target sounds occur at the beginning of word-initial stressless syllables, as in *[k]asino* and *[t]orpedo,* forming Condition 6. No additional conditions were identified for this environment.

The fourth aspiration environment is where target sounds occur at the beginning of a stressless syllable when immediately preceded by a stressless syllable and followed by a stressed one, as in words like *abra[k]adabra* and *lolla[p]alooza*, forming Condition 7. No additional conditions were identified for this environment.

Environment (Davis and Cho)	**Condition**	**Description**	**Example**
Environment 1	1	At the beginning of a syllable with primary stress; word-initial.	*[k]aravan*
	2	At the beginning of a syllable with primary stress; non-initial.	*re[p]eater*
Environment 2	3	At the beginning of a syllable with secondary stress; word-initial. **Three-syllable**	*[k]angaroo*
	4	At the beginning of a syllable with secondary stress; word-initial. **Five-syllable**	*[p]aradoxical*
	5	At the beginning of a syllable with secondary stress; non-initial.	*honey[k]omb*
Environment 3	6	At the beginning of a word-initial stressless syllable.	*[k]asino*
Environment 4	7	At the beginning of a stressless syllable when immediately preceded by a stressless syllable and followed by a stressed one.	*abra[k]adabra*

Table 1: Expected aspiration environments and production test conditions

3.1.2: Expected environments for unaspirated voiceless stops

The first environment where aspiration is not expected with voiceless stops is in target sounds in coda position at any stress level, as in *a[t]las* and *hy[p]nosis*, forming Condition 8. Note that this is an environment where not all selected test words are three-syllables, only five words were selected due to scarcity, and [t] sounds are expected to become glottal stops.

The second environment for unaspirated voiceless stops is at the beginning of a stressless syllable following a stressed syllable, as in *nau[t]ical* and *ra[p]id*, forming Condition 9. This is another environment where the word list was not able to be limited to three-syllable words. Here, [t] sounds are expected to flap.

The third environment for unaspirated voiceless stops is when the target sounds occur as the possible second member of an onset, as in *s[p]ectator* and *e[k]xposure*. This environment has been split into four conditions based on stress level and word position: stressed initial onsets, as in *s[t]adium* and *s[k]ydiver*, form Condition 10; stressless initial onsets, as in *s[p]aghetti* and *s[k]edaddle*, form Condition 11; stressed non-initial onsets, as in *e[k]xposure* and *dis[k]overy*, form Condition 12; and stressless non-initial onsets, as in *ins[t]itute* and *bas[k]et*, form Condition 13. Only five words were included from each of these conditions with the intent that, should significant behavioral changes surface between them, a full ten-word list could be produced for both word position or stress level variables by combining the relevant two of these four categories: Conditions 10 and 12 could be regarded together in order to discuss this environment as an onset to a stress syllable, and Conditions 11 and 13 could be regarded together to discuss this environment as an onset to an stressless syllable.

The fourth (presumed) environment for unaspirated voiceless stops was not discussed by Davis and Cho, and does not seem to have been discussed in the literature on the realization of aspiration, but should not host aspiration given the foot-initial description of aspiration. It occurs in the second syllable of a word-final stress lapse, as in *harmoni[k]a* and *universi[t]y*, forming Condition 14. Here, [t] sounds were expected to flap.

Environment (Davis and Cho)	Condition	Description	Example
Environment 1	8	In coda position at **any** stress level.	*hy[p]nosis*
Environment 2	9	At the beginning of a stressless syllable following a stressed one.	*ra[p]id*
Environment 3	10	As the possible second member of an **initial** onset to a **stressed** syllable.	*s[t]adium*
	11	As the possible second member of an **initial** onset to a **stressless** syllable.	*s[t]aghetti*
	12	As the possible second member of a **non-initial** onset to a **stressed** syllable.	*dis[k]overy*
	13	As the possible second member of a **non-initial** onset to a **stressless** syllable.	*bas[k]et*
Novel environment	14	At the beginning of the second syllable of a stress lapse at the end of a word	*harmoni[k]a*

Table 2: Expected environments of unaspirated voiceless stops and production test conditions

From the full word list, I created 5 lists of 23 words, with 2 words from each 10-word category and one from each 5-word category. For each of these lists, the order of the words was randomized when placing them into a question and answer frame, and five slide decks were created for each list, with unique random orderings. Words within each group appeared within the same question and answer frame, but the ordering of the words in the deck was unique for each randomized version. This process produced 25 unique slide decks, with 5 different word groups, numbered L1–L5, each with five different random slide orderings, labeled R1–R5.

A full list of all 23 question and answer frames is provided in Appendix B.

3.3 Procedure

This study was a self-paced reading task where participants read the 23 question and answer pairs from a slide on a computer, representing 1/5th of the complete word list. Participants were seated in a sound attenuated booth and wore a Shure WH30 condenser headset microphone recorded into a TASCAM DR-100 portable audio recorder. Participants were instructed to speak naturally, as if talking to a friend, to speak the entire phrase again if they mispronounced a word, and were asked to read through all of the slides before beginning recording to ensure that all words were familiar. After the conclusion of the study, the purpose and function of the study was explained to the participants.

3.4 Data cleaning

The data from three additional participants were excluded from the study: One for a non-functioning recorder, another due to a low recording volume that made accurately segmenting the sound file impossible, and a third due to hyperarticulation on test words, characterized by pauses between words and upwards intonation on target words. Incorrectly pronounced target words from the remaining 25 participants were excluded from the data in cases where the incorrect pronunciation occurred in a syllable adjacent to the target syllable or if it resulted in an added or dropped syllable, possibly affecting the VOT measurement. Cases where a pronunciation was not corrected but the mispronunciation was more than a syllable after the target syllable (such as pronouncing a different vowel) were included in the data. If participants corrected a pronunciation, the corrected pronunciation was always the one measured.

VOT durations were measured using the program Praat (Boersma & Weenick 2019).

3.4 Results

The data from this study show two clear groupings based on VOT duration. The majority of the expected aspiration conditions form a group that is well above the other conditions, displaying the behavior that would be expected of environments that are categorically aspirated. Most of the non-aspiration conditions form a group well below the other conditions, displaying the behavior that would be expected of environments that are not categorically aspirated.

The z-scored duration of VOT across all conditions is shown in Figure 5 below, separated by consonant. The z-scores show the change above or below each consonant's average VOT duration in terms of standard deviation from the mean, normalizing the variation between each consonant and speaker and thereby accounting for the speaker variation in VOT length that is discussed by Chodroff and Wilson (2017). There are several clear groupings across all three consonants; in Conditions 1-5, there is consistently a robust amount of VOT aligning with the expectation of categorical aspiration. In Conditions 8-13 there is a clearly much lower VOT, aligning with the expectation that these conditions are not categorically aspirated. Three conditions behave somewhat in between these two larger groupings; Conditions 6, 7, and 14. The VOT levels of the target consonant [p] and [k] for all three ambiguous conditions are significantly below the length of Conditions 1-5 and above the length of Conditions 8-13.

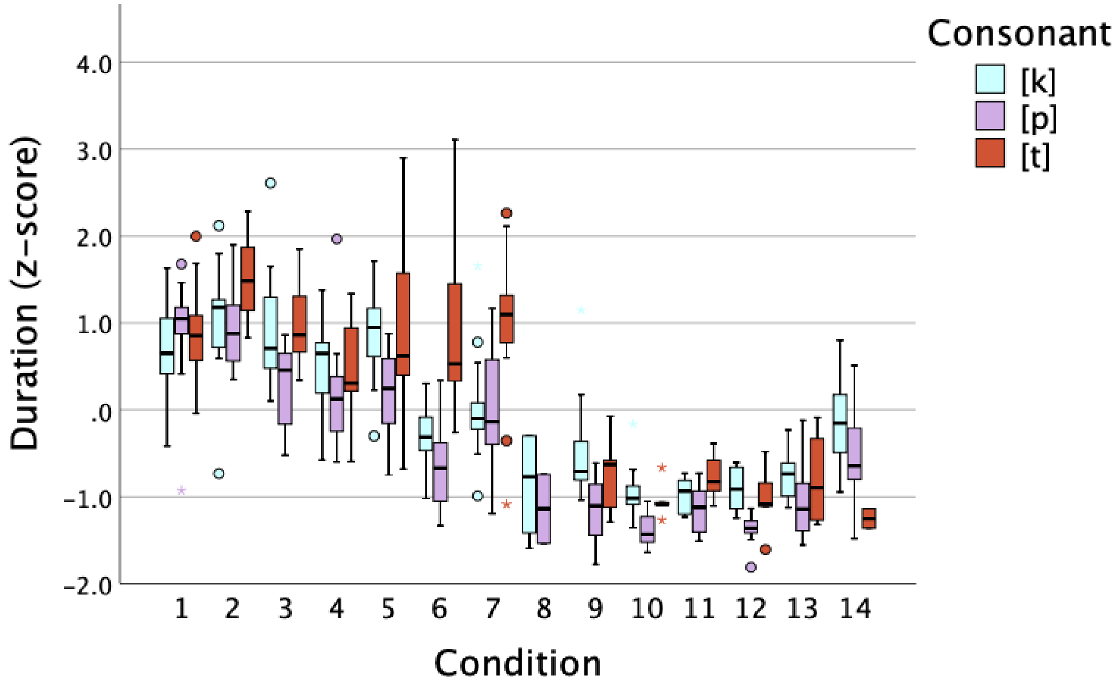

Figure 5: Z-score of voice onset time (VOT) duration by condition for consonants [k], [p], and [t]

The distribution of VOT for [k] in Figure 5 shows three groups of categorically aspirated, unaspirated, and ambiguous, with Conditions 6, 7, and 14 behaving ambiguously. The distribution of VOT for [p] is largely similar to that of [k], with primary stress aspiration conditions (1 and 2) being notably more aspirated than others, yet still showing a clear division from the unaspirated Conditions 8-13. Once again, Conditions 6, 7, and 14 are ambiguous in their categorical aspiration status. The distribution of VOT for [t] is distinct from that of [p] or [k]. In the case of [t], all of the expected aspiration conditions, including 7, behave similarly, with robust levels of VOT, and are very distinct from the unaspirated conditions and none of the test conditions behave ambiguously. However, [t] test words were intentionally underrepresented in the word list, and some conditions have insufficient numbers of measurements.

Note that the number of VOT measurements recorded in Condition 8 is limited: of the 10 VOT measurements for /k/, *a[k]ne* was pronounced with a glottal stop [ʔ] two out of five times and *a[k]tress* with a glottal stop four out of five times; of the 10 VOT measurements for /p/, *la[p]se* was pronounced with a glottal stop [ʔ] four out of five times and *hy[p]nosis* with a glottal stop four out of five times; and of the five VOT measurements for /t/, no words were pronounced with a [t], all surfacing with a glottal stop [ʔ]. The level of VOT recorded for Condition 8 in all consonants is thus based on six pronunciations rather than fifteen, and [t] is not present in the data at all.

The number of VOT measurements recorded in Conditions 9 and 14 for [t] is also limited. For Condition 9, of the 15 possible VOT measurements for the three words with [t] in the target position, *nau[t]ical* was pronounced with a voiced alveolar flap [ɾ] four out of five times, *en[t]ertain* with a flap once out of five, and *rhe[t]oric* with a flap in all five pronunciations. Condition 9 is thus based on five VOT measurements of [t] rather than fifteen. For Condition 14, of the 15 possible VOT measurements for [t], *personali[t]y* was pronounced with a flap four out of five times, *universi[t]y* with a flap three out of five times, and *electrici[t]y* with a flap in all five pronunciations. Condition 14 is thus based on three VOT measurements for [t] rather than fifteen. The lack of data for these conditions affirms the need to examine only [p] and [k], as in particular Condition 9 is the only syllable-initial unaspirated stop environment and

Condition 14 is the stress environment counterpart for Condition 7, both of which are critical comparison cases for the aspiration conditions.

A full list of the measured consonants in the data is provided in Table 3 below. Allophonic variants of [t] are not included. Conditions that had a five-stimuli word list are marked with an asterisk.

Condition	[k]	[p]	[t]	Total
1	20	14	15	49
2	15	19	15	49
3	20	15	14	49
4	15	20	15	50
5	25[1]	10	15	50
6	20	15	15	50
7	14	20	13	47
8*	4	2	0	6
9	15	20	5	40
10*	10	10	5	25
11*	10	10	5	25
12*	10	10	5	25
13*	10	10	4	24
14	20	15	2	37
TOTAL	207	188	131	526
% pronounced[2]	98.6%	91.7%	81.9%	91.5%

Table 3: Production study total recorded consonant count by condition, excluding allophonic stop variants

As expected, /t/ frequently did not surface as the voiceless stop [t], but instead surfaced as one of its allophonic variants, the voiced alveolar flap [ɾ] and the glottal stop [ʔ]. Despite the fact that it behaves differently in some conditions from [k] and [p], because its behavior cannot be analyzed relative to the other two voiceless stops in Conditions 8, 9, and (crucially) 14 I set aside data from [t] going forward.

Collapsed z-score measurements of VOTs in the recordings of both [p] and [k] pronunciations from the production study are shown in Figure 6 below. We see a clear group of conditions where aspiration is expected (Conditions 1-5) and a clear group of conditions where aspiration is not expected

[1] One additional [k] word was mistakenly included in the Condition 5 word list instead of a [p] word, resulting in 5 [k] words, 2 [p] words, and 3 [t] words.

[2] Based on the recorded number of pronunciations for each consonant divided by the total number of words with the target consonant in the word list multiplied by 5 (the number of speakers for each word list). 210 [k], 205 [p], and 160 [t] pronunciations were expected in the absence of allophonic variation and mispronunciation.

(8-13). These VOT durations largely align with the left-edge aspiration hypothesis: there is a clear division between most of the conditions in which aspiration was expected and most of the conditions in which it was not. There are, however, a group of outliers from this trend, conditions where the raw duration of VOT recorded for both [p] and [k] was somewhere in between the categorically aspirated conditions with long VOT measurements and the categorically unaspirated group with short VOT measurements, as seen in Figure 6. These outliers consisted of the initial stressless condition (*[k]abana,* Condition 6), the medial stress lapse condition (*lolla[p]alooza,* 7), and the final stress lapse condition (*Harmoni[k]a,* 14). While Condition 14 is not predicted to be aspirated as a rightward foot adjunct, its comparability with the supposed aspirated Conditions 6 (both stressless syllables) and 7 (both conditions in a lapse position) suggest it may be behaving in a similar manner to these two conditions. This group of three conditions might be understood together as showing the realization of aspiration in stressless syllables, where we find VOT to be somewhere in between the categorically aspirated positions and the unaspirated positions.

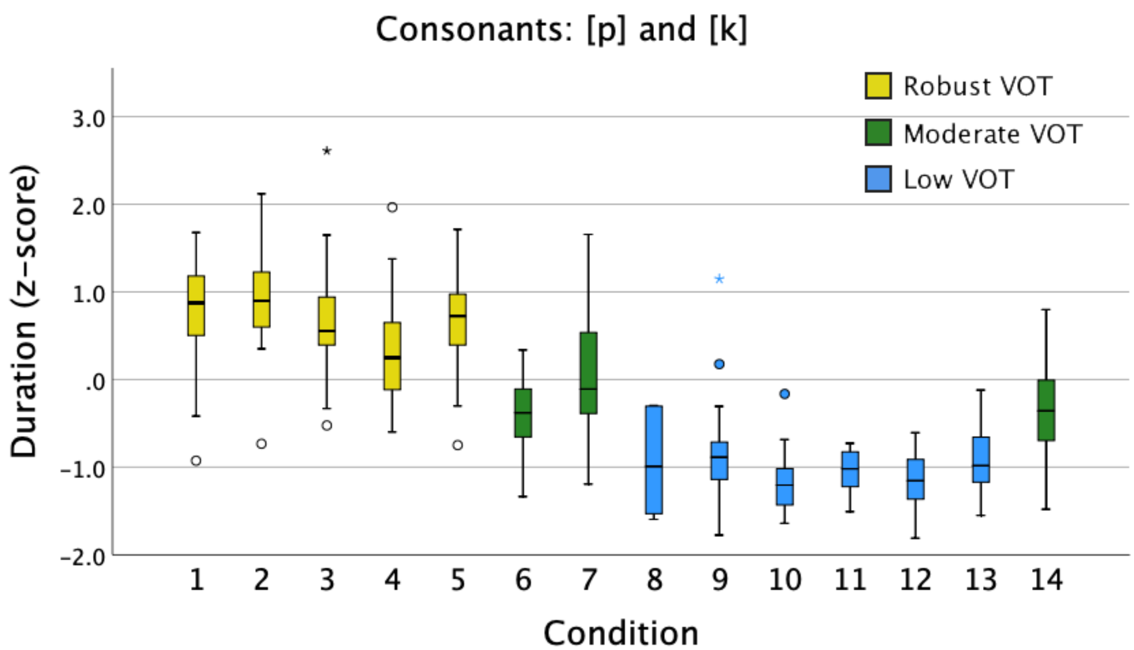

Figure 6: VOT duration by condition in the production study, divided by aspiration group of robust, moderate, or low

Looking at each individual participant's distribution of VOT across conditions, for each of Conditions 6, 7, and 14 the majority of participants (14-15 out of the 25 total) pronounced a moderate level of VOT, between the categorical groupings of aspirated and unaspirated. A handful (5-6) pronounced Conditions 6 and/or 14 with a low level of VOT that was like the level of their unaspirated conditions; this was never the case for Condition 7, which was pronounced with a more robust amount of VOT (8 participants) if it patterned with either grouping.[3]

A Generalized Linear Model (GLiM) was run for dependent variable (DV) *duration (z-score)* and factors *condition*, *consonant*, and their interaction. Conditions 10, 11, 12, and 13 were combined because they are in effect the same environment split into multiple groups for the purpose of comparison.

[3] Number of cases reported are based on a visual inspection of a plot of each individual's VOTs by condition (as in Figure 6). Three participants' conditions were not categorized because their aspirated and/or unaspirated conditions varied so greatly that the behavior of Conditions 6, 7, and 14 were not clear.

	Wald Chi-Square	*df*	Significance
(Intercept)	3.438	1	p=0.064
Condition	921.340	11	p<0.001
Consonant	31.814	1	p<0.001
Condition * Consonant	28.677	10	p=0.001

Table 4: Fixed effects chart for production generalized linear model

This model shows that there is a statistically significant interaction between *condition* and *consonant*; in some cases the VOT for [p] is higher than [k], and in some cases [p] is lower. However, I continue to collapse the two consonants going forward since this study is not analyzing the behavior of specific consonants, but rather aspirated consonants as a group, and there are roughly equivalent numbers of each consonant in all conditions. Collapsing the pool of available data significantly increases the sample size while also allowing us to examine the behavior of aspiration as a phonological feature rather than the behavior of aspiration on a specific consonant.

Examining the pairwise comparisons of the conditions finds statistically significant differences between pairs of aspirated conditions as well as pairs for aspirated and unaspirated conditions, meaning that we cannot use pairwise comparisons as a way to examine threshold effects between the expected aspirated and expected unaspirated groups. While statistical models are sensitive to absolutes, aspiration is known to be a categorical threshold effect, and I thus rely on visual comparison (as in Figure 6) rather than statistical comparison between conditions.

The existence of a visual group of conditions for the consonants [p] and [k] in which the presence of aspiration is unclear, in contravention of our understanding of aspiration as fundamentally categorical in English, motivates further exploration into how aspiration surfaces and is perceived in these conditions. While the phonetic realization of aspiration in these conditions is somewhat unclear in the way that they are produced, approaching them from a perceptual standpoint, i.e. focusing instead on how they are heard and interpreted, should allow us to disambiguate whether they are functioning more as aspirated or unaspirated syllables.

4.0 Experiments 2 and 3: Perception Studies

In order to disambiguate the aspiration status of the three marginal conditions from the production experiment, Conditions 6, 7, and 14, I deployed two consecutive perception studies. These studies seek to gauge whether the presence of aspiration in these positions is important for the listener in identifying the nature of the test words. If changes in VOT in these positions are perceptually salient to speakers of English, then aspiration is likely categorically encoded for these positions even if the acoustic realization is somewhat weaker than would be expected relative to what is recorded in other conditions. However, if changes in VOT are not salient to listeners, then the presence of aspiration in these positions does not significantly impact the identification of words, and it is likely not encoded in these conditions. I approach the question of perceptual salience from two slightly different perspectives in the two perception studies.

<u>4.1 Stimuli components</u>

Stimuli for both of the perception studies were created from the recordings generated by participants in the production study. These stimuli consist of the target word with the level of VOT of the target sound altered to four different levels: 15ms, 35ms, 55ms, and 75. These levels allow for comparison of participant response at different intervals, both above and below the perceptually important level of 35ms as the dividing line between aspirated and unaspirated, and also matching the high levels of VOT recorded in some aspirated words in the production study. These four levels allow us to observe a large (40ms) interval both across the 35ms threshold (by comparing 15ms and 55ms) and within crossing it (by comparing 35ms and 75ms), as well as examine small changes both below (15ms and 35ms) and above (55ms and 75ms) the threshold.

For consistency, the recordings of only one speaker per word list was selected to be used as the stimuli for the perception studies. I selected the first subject recording each word list, with certain exceptions. Recordings were rejected if they had no data for a condition that would be used in the perception study (such as a mispronounced word) or if the level of VOT recorded on a perception study condition was a statistical outlier, which could possibly influence the pronunciation of other parts of the word. Table 5 below provides the demographic information of the participants whose production recordings were used for the perception study stimuli.

List	Participant no.	Gender	Age
1	11	M	25
2	12	M	21
3	34	F	19
4	20	F	18
5	31	M	19

Table 5: Demographic information of perception study stimuli speakers

Data from [t] pronunciations were not included in the stimuli list for either perception experiment because of the tendency of [t] to surface as allophonic variants in certain environments and its intentional underrepresentation in the word list.

Based on the results of the production study, I selected 9 conditions to be tested in the perception studies: the seven expected aspiration conditions, one comparison unaspirated condition (*ra[p]id*, 9), and the ambiguous final lapse condition (*Ameri[k]a*, 14). Condition 9 is the only unaspirated condition other than 14 in which the target consonant is in the onset of a syllable without being part of a consonant cluster, making the VOT measurable, though its location in the onset of a syllable immediately following a stressed syllable means that it is the second member of a foot, and thus should not surface as aspirated. This condition also has a full word list of 10 tokens, in parallel to Conditions 1-7 and 14. This provides the best comparison case for a categorically unaspirated voiceless stop.

<u>4.2 Experiment 2: Categorical perception</u>

The categorical perception study is focused on listeners' ability to recognize phonological aspiration in the absence of any external frame of reference or comparison. In this study, participants were presented with a VOT-modified pronunciation of a word and responded with a judgment based on whether they perceived a target syllable as voiced or not. Through this study, I hope to determine whether different conditions have distinct thresholds for what is considered categorically aspirated by listeners, and what factors might influence any differences in threshold that might exist.

<u>4.2.1 Participants</u>

There were 40 participants for the categorical perception study, 20 taking the A version of the study and 20 taking the B version. Each version had 120 stimuli. Participants received course research credit for participating. The purpose of the study was explained to participants, and they were given the opportunity to withdraw at any time.

Of the 40 participants, 31 identified themselves as female and 9 as male. The mean age of participants was 19, with an age range of 18-23 years.

4.2.2 Stimuli

Stimuli consisted of the modified VOT tokens described in §4.1. There were two versions of the study, each consisting of 120 stimuli played individually, including two pronunciations of each of the 60 test words. Each version included two VOT levels out of the four possible levels for each word, with VOT differences of 40ms between the two levels present in the same version (eg. 15/55 or 35/75). Each study version included an equal balance of all four VOT levels within each condition.

4.2.3 Procedure

Participants heard stimuli and were presented with two buttons representing the standard spelling of the word and a spelling corresponding to a voiced stop at the target position (e.g. *cabana* or *gabana*). The version with unconventional spelling was always presented on the left so that participants reading left to right always encountered the altered spelling first in order to maximize participants' awareness of the alternative voicing. Participants were told that the stimuli included some pronunciations that had been altered to match the unconventional spelling on the screen. Participants were asked to click on the spelling that they heard. There was a self-timed break half-way through the experiment.

A full list of alternative spellings for categorical perception stimuli is provided in Appendix C.

4.2.4 Results

The identification of the altered stimuli versions aligned almost entirely with the expected distribution of categorical aspiration. In conditions where strong aspiration was present in the original recordings (Conditions 1-5), listeners were much more sensitive to the level of VOT present, and consistently identified the 15ms level as the word's voiced equivalent (eg. identifying the 15ms level of *consonant* as *gonsonant*). This also occurred for the initial stressless condition (*[p]otato,* 6), which had been ambiguous in the production study, indicating that aspiration is categorically present in this condition despite its less forceful acoustic realization in the production study. The other three environments in the study behaved as a group: medial stress lapse (*lolla[p]alooza,* 7), onset of weak member of a foot (*ra[k]quet,* 9), and final stress lapse (*Ameri[k]a,* 14). All three were consistently identified as being the altered spelling at a much lower rate, despite having the same 15ms level of VOT as the other conditions that participants much more easily identified as altered.

Figure 7 shows the likelihood of a word in each condition being identified as its voiced equivalent over the four VOT lengths (15ms, 35ms, 55ms, and 75ms). A higher percentage of alternative word identification indicates a more frequent response of the voiced spelling for a given stimulus. Above the categorical aspiration threshold at 35ms VOT and above there are no groupings of condition behavior, indicating that in all conditions words are being perceived as voiceless. There is a clear split in the 15ms group between a categorically aspirated set of conditions (1-6) and a categorically unaspirated set of conditions (7, 9, 14).

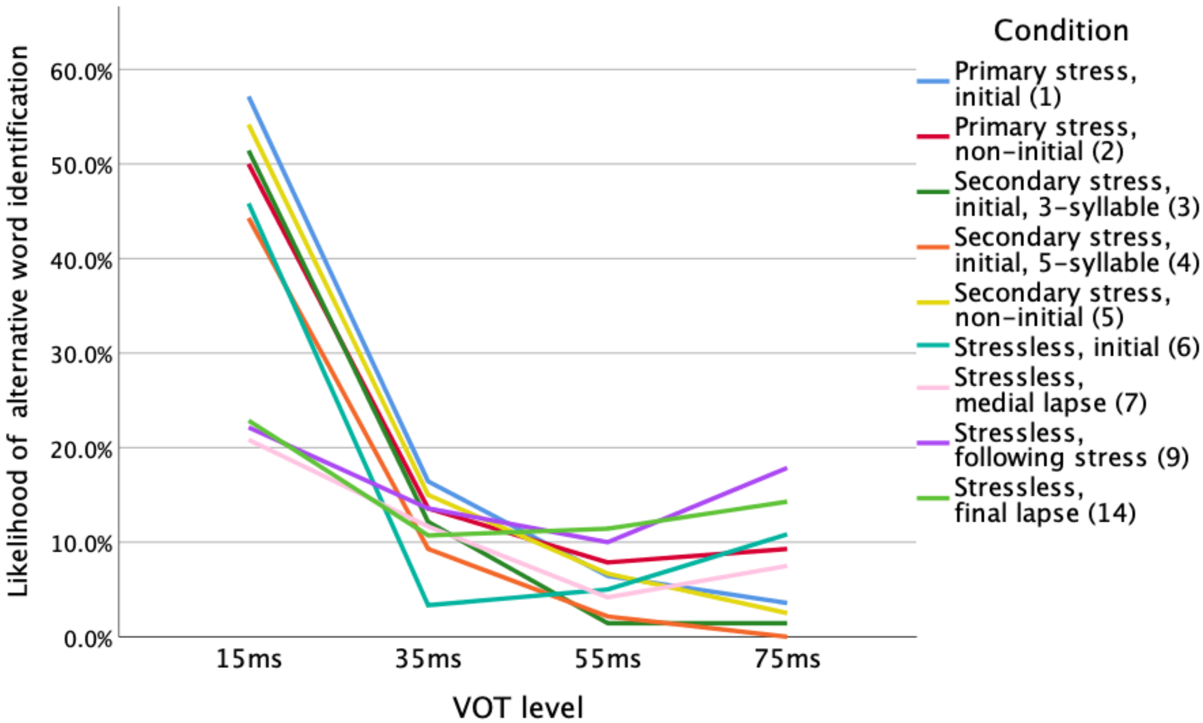

Figure 7: Categorical perception study likelihood of alternative (voiced) word identification by stimuli VOT level.

A binary logistic Generalized Linear Mixed Model (GLMM) was run for dependent variable (DV) *response* and factors *condition*, *length*, and their interaction, with individual intercepts fitted for the random effects of *subject* and *word*. We see that the interaction of *condition* and *length* is significant.

Source	F	*df1*	*df2*	Sig.
Corrected model	18.264	17	176	p<0.001
Condition	2.668	8	52	p=0.016
Length	236.512	1	2382	p<0.001
Condition * Length	5.825	8	2382	p<0.001

Table 6: Fixed effects chart for categorical perception GLMM

We are interested in the two groupings we visually see in the graph in Figure 7 at 15ms. Pairwise comparisons between the conditions at 15ms establish that Conditions 1-6 are not statistically significantly different from each other ($p \geq 0.109$) and Conditions 7-9 are not statistically different from each other ($p \geq 0.799$). All the conditions within the two groups are statistically significantly different ($p \leq 0.008$).

We can also look at pairwise comparisons within each condition between 15ms and 35ms, assessing whether there is a significant change in behavior between those two VOT lengths. As we see in Table 7 below, there is a significant difference in the behavior of Conditions 1-6 between 15ms and 35, to be expected given their categorical aspiration. There is a somewhat marginal difference between the two VOT levels for Conditions 7 and 9. Condition 14, conversely, has a quite robust difference between the two VOT levels, with a much greater significance than 7 and 9, suggesting a more robust perceptual distinction between 15ms and 35ms. It is somewhat surprising that Condition 14 shows such a difference. However, since the identification of the alternate (voiced) version of the word is so low at 15ms (around 23%) we can take this difference between the 15ms and 35ms VOT levels to be a gradient effect (as it does not cross the aspiration threshold). Responses for both 15ms and 35ms VOT levels occur below the amount of VOT at which listeners hear a voiced stop, even in the absence of stop voicing in the acoustic signal.

Condition	Significance between 15ms and 35ms
1	p<0.001
2	p<0.001
3	p<0.001
4	p<0.001
5	p<0.001
6	p<0.001
7	p=0.051
9	p=0.044
14	p=0.007

Table 7: Pairwise comparisons of categorical perception conditions between 15ms and 35ms.

This data shows a split in behavior from the three environments that showed moderate levels of VOT for [p] and [k] words in the production study. Condition 6 behaves like other aspirated environments (1-5) while Conditions 7 and 14 behave like Condition 9, the non-aspiration environment included as a control. We see a clear grouping of categorically aspirated syllables (Conditions 1-6) and categorically unaspirated syllables (Conditions 7, 9, and 14), with both stress lapse environments unambiguously perceived as unaspirated and patterning together despite their supposed different positions in relation to foot boundaries.

Note that the percentage of words identified as the alternative version peak around 55% for Condition 1. While there is a clear split between the categorically aspirated conditions and Conditions 7, 9, and 14, the reasonably low percentage score for the 15ms aspirated conditions relative to the possible 100% score is likely due to top-down cues. Participants are heavily influenced by the presence of the voiceless versions of test words in their mental grammars, significantly depressing the identification of non-word voiced versions, as described in Ganong (1980). In short, while there is a word *positive* in participants' mental grammar, there is no voiced equivalent *bositive*, and thus they are heavily biased towards hearing *positive* regardless of the actual VOT level heard. Despite this effect, there remains a clear and significant distinction between the behavior of categorically aspirated conditions and unaspirated conditions in this study.

4.3 Experiment 3: Gradient Perception

The gradient perception study is intended to probe the perception of aspiration through a same/different task, and is focused on participants' sensitivity to VOT change at target locations in stimuli words. The study presents pairs of stimuli with various relationships to the threshold, and seeks to determine the importance of a threshold-based perception of VOT in different environments. This should reveal whether categorical aspiration is a perceptually important cue to listeners in these conditions, regardless of whether it is produced by speakers.

4.3.1 Participants

There were 60 participants for the gradient perception experiment, 30 taking the A version and 30 taking the B version. Each participant had 180 stimuli. Participants received course research credit for participating. The purpose of the study was explained to participants, and they were given the opportunity to withdraw at any time.

Of the 60 participants, 39 identified themselves as female, 20 as male, and 1 as non-binary. The mean age of participants was 19, with an age range of 18-26 years.

4.3.2 Stimuli

Stimuli consisted of the modified VOT tokens described in §4.1. There were two versions of the study, A and B, each consisting of 180 stimuli played in pairs. Pairs consisted of two modified pronunciations of the same word, each word being presented three times: once in a pair with identical VOT lengths (either 15ms/15ms or 55ms/55ms), once with VOT lengths contrasting across the 35ms threshold (either 15ms/35ms or 15ms/55ms), and once with VOT lengths contrasting without crossing the 35ms threshold (either 35ms/55ms or 35ms/75ms). The ordering of all three pairs were reversed in equal numbers, so that the longer VOT stimuli appeared first and second in equal numbers. Each study version included a balance of identical pair VOT levels and of contrastive VOT orderings. Stimuli were separated by a silent 100ms inter-stimulus interval.

Stimuli pairings were broken down by type, where Type 1 and 2 pairings were identical VOT pairs, Type 1 at 15ms and Type 2 at 55ms, Type 3 and 4 pairings were VOT distinctions crossing the aspiration threshold, Type 3 with a 20ms difference and Type 4 with a 40ms difference, and Type 5 and 6 pairings were VOT distinctions not crossing the threshold, Type 5 with a 20ms difference and Type 6 with a 40ms difference. These Types are described in Table 8 below.

Type	VOT level distinction	Description
1	15ms / 15ms	Same VOT level below the typical aspiration threshold
2	55ms / 55ms	Same VOT level above the typical aspiration threshold
3	15ms / 35ms 35ms / 15ms	Small (20ms) VOT difference crossing the aspiration threshold
4	15ms / 55ms 55ms / 15ms	Large (40ms) VOT difference crossing the aspiration threshold
5	35ms / 55ms 55ms / 35ms	Small (20ms) VOT difference not crossing the aspiration threshold
6	35ms / 75ms 75ms / 35ms	Large (40ms) VOT difference not crossing the aspiration threshold

Table 8: Gradient perception study stimuli pairing types

Participants should consistently identify Types 4 and 6, which have 40ms acoustic differences, as different more frequently than they identify Types 3 and 5, which have 20ms acoustic differences, as different. In conditions where aspiration is an important cue to word identity, differences in VOT that cross the threshold (Types 3 and 4) should be identified as different more frequently than their equivalent length distinction (Types 5 and 6) are identified as different. In conditions where aspiration is not an important cue, we do not expect to see significant differences in behavior between the threshold-crossing and non-threshold-crossing stimuli, but only based on VOT length difference.

<u>4.3.3 Procedure</u>

Participants were primed by hearing two sample pairs of stimuli: *[k]onduit* (Condition 1) with a Type 4 distinction (15ms/55ms) and *pro[p]ulsion* (Condition 2) with a Type 6 distinction (35ms/75ms). These priming stimuli were selected for clearly presenting changes in aspiration both word-initially and word-internally on clearly aspirated syllables (primary stress locations), and with VOT distinctions of the maximal acoustic difference of 40ms, one crossing the threshold and the other not. The two priming words came from different word lists with different speakers, exposing participants to two different speakers during the priming phase. The waveform of the first priming stimulus pair (*[k]onduit,* 15ms/55ms) is shown in Figure 8 below, highlighting the VOT difference.

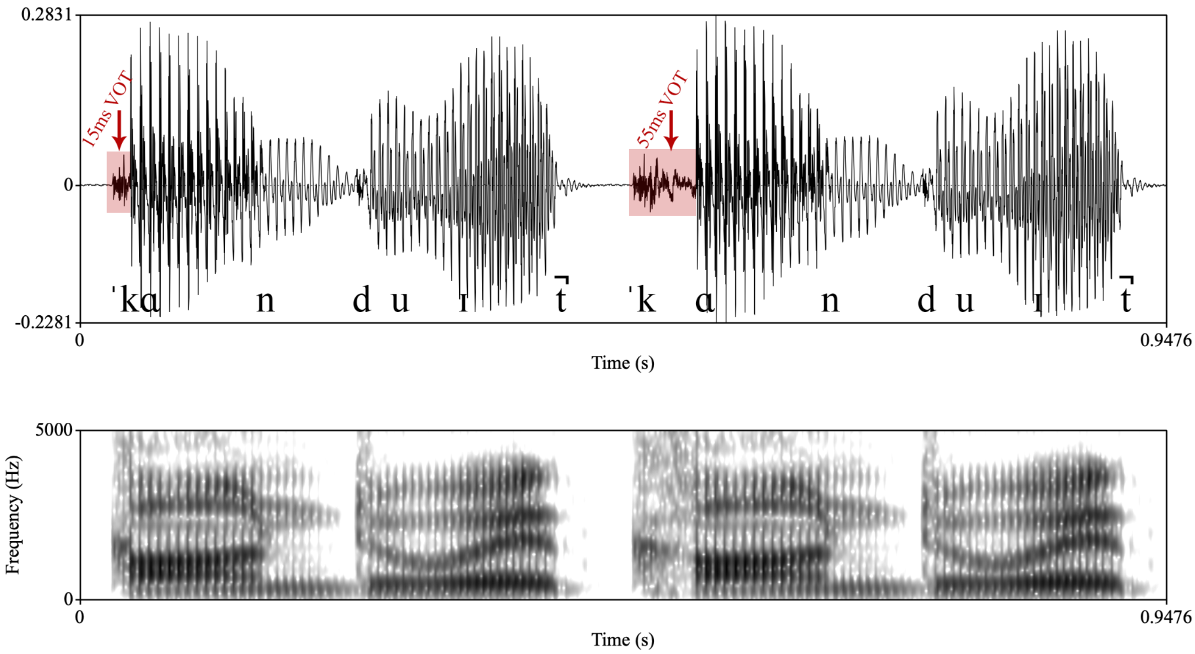

Figure 8: Gradient perception study first priming stimulus ([k]onduit, Condition 1) with 15ms vs 55ms VOT difference highlighted.

Participants were played each priming stimuli twice, and were specifically directed to the difference. They were told that differences between the two pronunciations, if present, would manifest as a difference in the strength of a ‘p’ or ‘k’ sound somewhere in the word.

During the study, participants heard pairs of stimuli corresponding to the six types discussed in §4.3.2 above. Participants were instructed to go with their instinct if not sure whether the pronunciations in a pair were the same or not. Responses were collected by keypress, where ‘f’ indicated the same pronunciation and ‘j’ indicated a slight difference between the two pronunciations. There was a self-timed break halfway through the study.

4.3.4 Data cleaning

The data from 30 additional participants was excluded from the study results. The criteria for inclusion was that subjects had to respond “different” more to Type 4 (15ms/55ms) pairings than to Type 2 (55ms/55ms) pairings by a margin of 3 judgements (correlating to 10% of the 30 stimuli pairs of each type). This excluded participants who failed to identify an identical stimuli pairing (Type 2) as being the same reasonably more frequently than the maximally different stimuli pairing (Type 4), which had both a 40ms difference in VOT between stimuli and crossed the aspiration threshold.

4.3.5 Results

As a baseline metric for the effectiveness of the gradient perception task, I first compare the responses to 40ms VOT difference types (Types 4 and 6) with 20ms difference types (Types 3 and 5). If participants are noticing and reacting to the VOT level changes between the two stimuli, then the 20ms types (3 and 5) should both have a higher “same” response rate than the 40ms types (4 and 6). Table 9 below compares the percentage of “same” responses for each type, showing that there is a notable reduction in “same” responses between 40ms VOT change types and 20ms VOT change types.

Type	Percent of pairs identified as the same
Type 3 (15ms,35ms) 20ms crossing threshold	65%
Type 4 (15ms,55ms) 40ms crossing threshold	48%
Type 5 (35ms,55ms) 20ms no crossing threshold	69%
Type 6 (35ms,75ms) 40ms not crossing threshold	56%

Table 9: Comparison between 40ms VOT change types and 20ms VOT change types

With evidence that participants are able to consistently identify 40ms changes in VOT as “different” more than 20ms changes in VOT, I now examine how participants respond to different types that have the same level of VOT yet cross the threshold for aspiration.

A binary logistic Generalized Linear Mixed Model (GLMM) was run for dependent variable (DV) *response* and factors *condition*, *type*, and their interaction, with individual intercepts fitted for the random effects of *subject* and *word*. We see that the interaction of condition and type is significant.

Source	F	*df1*	*df2*	Sig.
Corrected model	6.493	47	326	p<0.001
Condition	.919	7	37	p=0.503
Type	45.276	5	460	p<0.001
Condition * Type	2.722	35	810	p<0.001

Table 10: Fixed effects chart for gradient perception GLMM

As discussed above, there is a significant difference (p<0.001) between the behavior of both 40ms VOT difference types and both 20ms VOT difference types, both crossing and not crossing the aspiration threshold, indicating that participants are significantly more likely to respond “different” to a larger VOT change than a smaller change.

We are particularly interested in the comparison of Type 3 to Type 5, as both have 20ms differences within the pair, but Type 3 is across the aspiration threshold (of ~30ms) while Type 5 is not. Looking at the pairwise comparisons for *condition*type*, only Condition 5 showed a statistically significant difference between Types 3 and 5. The second comparison of interest is between that of Type 4 to Type 6, which both have a 40ms difference, with the former being across the aspiration threshold and the latter not. Again looking within each condition, we see that all the known aspiration environments show a statistically significant difference. Condition 9, however, does not. Our test cases are split, where Condition 7 behaves as the aspirated ones do, and Condition 14 does not. Both Type 3 to 5 and Type 4 to 6 significance comparisons are shown in Table 11 below.

Condition	Type 3 and Type 5 (20ms) significantly different?	Type 4 and Type 6 (40ms) significantly different?
1	No (p=0.396)	Yes (p=0.002)
2	No (p=1.000)	Yes (p=0.008)
3	No (p=0.326)	Yes (p<0.001)
4	No (p=0.170)	Yes (p=0.032)
5	Yes (p=0.018)	Yes (p=0.014)
6	No (p=0.208)	No (p=0.742)
7	No (p=1.000)	Yes (p=0.012)
9	No (p=0.292)	No (p=0.072)
14	No (p=0.154)	No (p=1.000)

Table 11: Comparison of significance between Type 3 and 5 pairing and Type 4 and 6 pairing by condition. All p-values are multiplied by 2 for the Bonferroni correction.

The identification of stimuli pairs as “different” when crossing the aspiration threshold largely aligned with the expected aspiration status of each condition. When asked to differentiate between a pair of stimuli with a difference of 40ms whose VOT levels crossed the threshold for categorical aspiration, participants consistently responded that pairs were different in Conditions 1,2, 3, 4, 5, and 7, indicating that categorical aspiration is an important cue in these conditions. Participants were not better at noticing a distinction when stimuli pairs crossed the categorical aspiration threshold in Condition 9, the control, providing an example of what a categorical lack of aspiration looks like perceptually. Participants were also unable to mark a difference between threshold-crossing and non-threshold-crossing stimuli pairs in Conditions 6 and 14, suggesting that neither are categorically aspirated. For Condition 6, this is in opposition to the behavior observed in the categorical perception study.

A comparison of participants’ likelihood of returning a “same” response for each of the conditions in the two 40ms-difference stimuli pairings (Type 4, 15ms / 55ms; and Type 6, 35ms / 75ms) is shown in Figure 9 below. Conditions that are categorically aspirated should see the 15ms/55ms stimuli pairing being identified as “same” markedly less than the 35ms/75ms stimuli pairing. Conditions that are not categorically aspirated should see no clear distinction between the two stimuli pairing types, as aspiration is not an important acoustic cue in these environments. Distinctions are seen in the difference between the two types rather than in the numerical proportion of “same” responses.

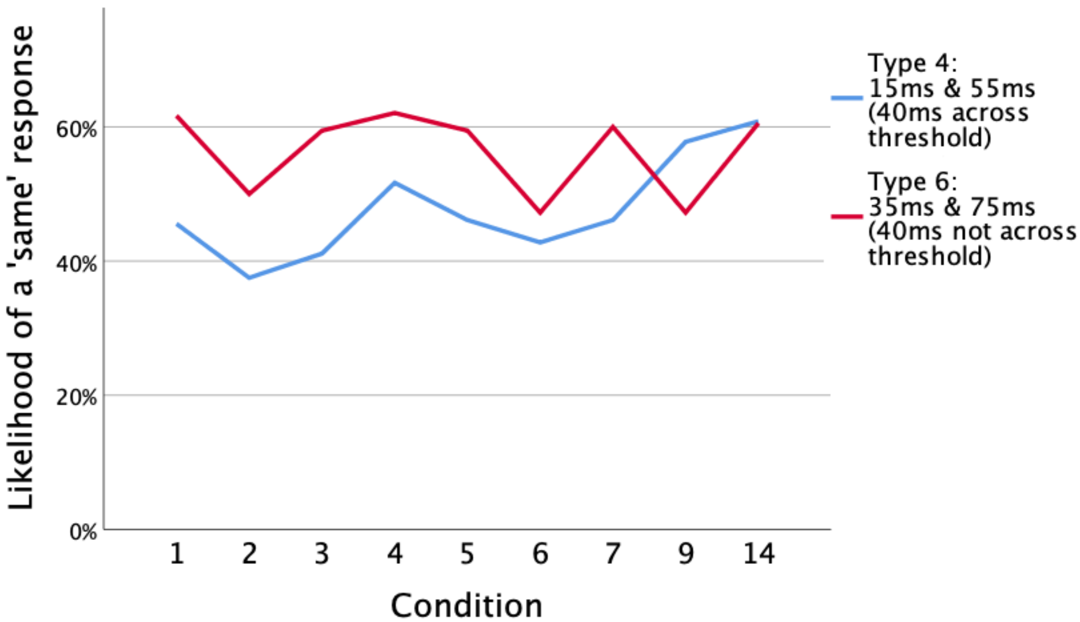

Figure 9: Gradient perception study likelihood of a “same” response by condition for 40ms-difference stimuli pairs.

In Figure 9 above, the responses for Types 4 and 6 in Conditions 1-5 and 7 are all distinctly different, with Type 4 (40ms across threshold) consistently being identified as “different” more than Type 6 (40ms not across threshold), showing that a threshold effect is important to participants in these conditions. In Conditions 6, 9, and 14, there is no clear difference between the behaviors of Type 4 and Type 6 stimuli pairs, suggesting that there is no important threshold effect in these conditions. For Condition 6, responses to Type 4 and Type 6 pairings have the same relation to each other as the categorically aspirated conditions, with Type 4 being identified as “same” more frequently than Type 6, but their difference from each other does not reach statistical significance.

5.0 Discussion

I deployed three experiments in this study on aspiration. In the production task, I examined how robustly aspiration is produced in all possible environments, and how the VOT duration in categorically aspirated conditions compares to that of categorically unaspirated conditions. These phonetic measurements resulted in three ambiguous conditions, with levels of VOT intermediate to the otherwise categorical distribution: the initial stressless environment (6, *[p]otato*), word-medial stress lapse environment (7, *lolla[p]alooza*), and the word-final stress lapse environment (14, *harmoni[k]a*). All other conditions behaved categorically as was expected.

In the categorical perception task (perception 1), I examined the importance of raw VOT duration change on the perception of voiceless stops, determining whether high VOT corresponds to voiceless identifications and low VOT corresponds to voiced identifications. Categorically aspirated conditions were identified as voiced when they occurred with a VOT of 15ms, while categorically unaspirated conditions were identified as voiceless at all VOT durations. Condition 6 patterned with other aspirated conditions, and Conditions 7 and 14 both patterned with the unaspirated control condition (9, *ra[p]id*).

In the gradient perception task (perception 2), I examined the sensitivity of listeners to VOT change in different conditions, determining whether the same amount of VOT change would be perceived differently depending on if it crossed the aspiration threshold or not. Categorically aspirated conditions exhibited a threshold effect, with a clear perceptual difference between threshold-crossing and non-threshold-crossing stimuli pairs. While Condition 6 exhibited the same general relationship between Types as the other aspirated conditions, this relationship did not reach statistical significance. Condition 7 clearly patterned with aspirated conditions and Condition 14 patterned with unaspirated Condition 9.

Based on the results of these three experiments, there are a number of conclusions that can be drawn concerning the behavior of the different conditions. For the most part, conditions that were expected to be aspirated consistently behaved as aspirated; Conditions 1-5 were robustly produced and perceived as aspirated in all three experiments. Likewise, most conditions that were not expected to exhibit aspiration behaved distinctly differently from the aspirated conditions; Conditions 8-13 behaved identically in the production task, and, through the representative behavior of Condition 9, were always quite distinct from aspirated conditions.

Condition 6 behaved ambiguously in the production study, with a lower level of VOT than would be expected given its presumed aspiration status, but its significant relationship to other aspirated conditions in the categorical perception study clarify its status as a categorically aspirated condition. In the gradient perception study Condition 6 again behaved strangely; there was no statistically significant difference between the 40ms VOT difference pairs that crossed or did not cross the aspiration threshold, and thus no threshold effect, indicating that it did not behave as aspirated in this study. However, the relationship between the two 40ms VOT difference pairs was the same as the other aspirated conditions, simply not to the same magnitude of difference, and without reaching statistical significance. The behavior of Condition 6 as aspirated in the categorical study and unaspirated in the gradient study seems to be an impossibility, as a threshold effect was clearly and significantly present in one study and not in the other.

Since the categorical perception study demonstrated that a threshold exists for Condition 6, we must assume that it is aspirated. In the gradient perception study, as visible in Figure 9, even in the most phonetically distinct and acoustically prominent aspiration conditions (1, [k]aravan, and 2, re[p]eater) the mean response hovers at around 40% “same.” Only ~60% of responses in these conditions are judgements that the highest amount of difference present in the study is indeed different. The distinctions between the stimuli in the gradient study are clearly perceptually very subtle, yet Condition 6 is still exhibiting the same general behavior as categorically aspirated Conditions 1-5 and 7, with Type 4 identified as “different” more than Type 6, but not to the level of statistical significance. Given the difficult nature of the study, perhaps more participants are needed in order to see if the same difference is consistently present, or if Condition 6 resolves into the same behavior as the other expected categorically aspirated conditions.

Initially in a similar manner to Condition 6, both Conditions 7 and 14 behaved ambiguously in the production study, with recorded VOT levels falling between the categorically aspirated and unaspirated groupings. For both conditions this is surprising; Condition 7 is expected to behave as categorically aspirated, and Condition 14 as unaspirated. To clarify this behavior, I examine their behavior in both perception studies relative to Condition 9, the most structurally related condition known decisively to be unaspirated, and to the behavior of the clearly aspirated Conditions 1-5. In the categorical perception study, neither Condition 7 nor Condition 14 behave as aspirated; the consistency of voiceless perception responses to both conditions even at the 15ms VOT level suggests that listeners identify both conditions as voiceless even without robust VOT, in parallel with the known unaspirated Condition 9. Low levels of VOT are tolerated without the consonant being heard as a voiced stop, suggesting that neither are categorically aspirated.

In contrast to the categorical perception study, the gradient perception study shows a significant aspiration threshold for Condition 7, and no such threshold for Condition 14. Participants were much better at distinguishing between two Condition 7 stimuli pronunciations when the difference in VOT crossed the threshold than when it did not, showing a clear threshold effect on par with those of aspirated Conditions 1-5. In Condition 14, no such threshold effect was observed, in parallel to Condition 9. In these conditions where no threshold effect was observed, the only difference perceived by participants was absolute difference in VOT, without regard to the aspiration threshold.

Condition 7 can thus be described as categorically aspirated. While in the production study it lacked a robust amount VOT relative to other aspirated conditions, and in the categorical study listeners tolerated a low level of VOT while still identifying it as a voiceless stop (like unaspirated Conditions 9 and 14), in the gradient perception study it robustly behaved as a voiceless stop, and exhibited the same significant threshold effect as other aspirated conditions.

This behavior could be explained as being an effect of the weak position in which the aspirated stop in Condition 7 surfaces. Unlike other aspiration locations, which are in stressed syllables or on the left edge of a word, aspiration in Condition 6 surfaces in the second consecutive unstressed syllable (stress lapse) in the middle of a word. This is a perceptually and phonetically weak position, and so it would thus not be surprising if speakers did not pronounce an acoustically strong feature such as aspiration as robustly in this position. Perceptually, this structural weakness could mean that listeners do not expect to hear as robust VOT in this position even though it is categorically aspirated, and therefore interpret stops here as voiceless regardless of the level of VOT produced and heard here. However, even though robust VOT is not produced or expected in this position, it can certainly be perceived, with listeners clearly able to differentiate stops above and below the aspiration threshold in this position.

The empirical evidence from this study thus shows that Conditions 1-7 all behave as categorically aspirated, and Conditions 8-14 all behave as categorically unaspirated. While Condition 7 is produced somewhat weaker than would be expected, it is perceived as aspirated, with a clear threshold effect, and is therefore a member of the class of categorically aspirated conditions. Though it had not been previously considered in literature on aspiration, Condition 14 behaves as categorically unaspirated in line with expectations. This provides clear support for the left-edge prosodic description of aspiration posited by Kiparsky (1979), Withgott (1982), and Jensen (2000), and affirms that all examined environments for aspiration can be accurately described under the prosodic definition.

This empirical support for the left-edge rule also seems to support Jensen's (2000) theory that English adjunction takes place first to the right and then to the left, resulting in the creation of left foot adjuncts (like that in Condition 6) prior to right foot adjuncts. In opposition to McCarthy's (1982) description of two equally acceptable infixation locations, this evidence suggests additional structural support for the infixation location shown in Figure 4a, where infixed expletives must precede the superfoot boundary, as in *lolla-fuckin-palooza*.

6.0 Conclusion

The purpose of this thesis was to examine the existing theoretical understanding of aspiration as appearing on the left edge of prosodic feet from a phonetic lens. I used a production study to measure the amount of voice onset time (VOT) in each possible environment for aspiration as well as environments for unaspirated voiceless stops in English. From this study, I concluded that aspiration largely did align with this left-edge description, although the initial stressless environment and the word-medial stress lapse environment did not have as robust a level of VOT as the other aspirated conditions, and the word-final stress lapse environment had a higher recorded level of VOT than would be expected from a categorically unaspirated environment.

Based on these results, I deployed two perception studies, one categorical and one gradient, to examine how listeners perceive VOT in relation to the threshold of aspiration. The categorical study revealed that the initial stressless environment was patterning with other aspirated conditions, while both lapse conditions were behaving like the unaspirated control condition. The gradient study revealed that the medial lapse condition robustly behaved like other aspirated conditions, with a strong threshold effect, while the final lapse condition behaved like the unaspirated control condition. Based on these results, I conclude that the medial lapse condition is categorically aspirated, and the final lapse is categorically unaspirated.

These phonetic and perceptual observations provide empirical support for the description of aspiration as occurring at the left edge of prosodic feet, including recursive feet, as described in Kiparsky (1979), Withgott (1982), and Jensen (2000), providing crucial qualitative support for this theory that had previously been lacking.

References

Abramson, Arthur S, and D.H Whalen. 2017. Voice Onset Time (VOT) at 50: Theoretical and Practical Issues in Measuring Voicing Distinctions. *Journal of Phonetics* 63: 75–86. https://doi.org/10.1016/j.wocn.2017.05.002

Boersma, Paul & David Weenink. 2019. Praat: Doing phonetics by computer [Computer program]. Version 6.1.47. Retrieved from http://www.praat.org/

Cho, Taehong, and Peter Ladefoged. 1999. Variation and universals in VOT: evidence from 18 languages. *Journal of Phonetics* 27:207-229. https://doi.org/10.1006/jpho.1999.0094

Chodroff, Eleanor and Colin Wilson. 2017. *Journal of Phonetics* 61:30-47. https://doi.org/10.1016/j.wocn.2017.01.001

Davis, Stuart and Mi-Hui Cho. 2003. The distribution of aspirated stops and /h/ in American English and Korean: An alignment approach with typological implications. *Linguistics* 41:607-652. https://doi.org/10.1515/ling.2003.020

Ganong, W.F., 1980. Phonetic categorization in auditory word perception. Journal of experimental psychology: Human perception and performance, 6(1), p.110.

Hammond, Michael. 2020. Prosodic phonology. *The handbook of English linguistics*: 365-384. https://doi.org/10.1002/9781119540618.ch19

Jiang, Jintao, Marcia Chen, and Abeer Alwan. 2006. On the perception of voicing in syllable-initial plosives in noise. *The Journal of the Acoustical Society of America* 119:1092-1105. https://doi.org/10.1121/1.2149841

Jensen, John. 2000. Against Ambisyllabicity. *Phonology* 17:187–235. http://www.jstor.org/stable/4420169

Johnson, Keith. 2012. *Acoustic and Auditory Phonetics*. 3rd ed. Wiley-Blackwell.

Kager, René and Violeta Martínez-Paricio. 2018. The internally layered foot in Dutch . *Linguistics* 56:69-114. https://doi.org/10.1515/ling-2017-0037

Kim, Sahynag, Jiseung Kim and Taehong Cho. 2018. Stop voicing contrast in American English: Data of individual speakers in trochaic and iambic words in different prosodic structural contexts. *Data in Brief* 21:980-988, https://doi.org/10.1016/j.dib.2018.10.053.

Kiparsky, Paul. 1979. Metrical Structure Assignment Is Cyclic. *Linguistic inquiry* 10: 421–441. https://www.jstor.org/stable/4178120

McCarthy, John J. 1982. Prosodic Structure and Expletive Infixation. *Language* 58: 574–90. https://doi.org/10.2307/413849.

Reetz, Henning and Allard Jongman. 2009. *Phonetics: Transcription, production, acoustics, and perception.* Wiley-Blackwell.

Repp, Bruno H. 1979. Relative Amplitude of Aspiration Noise as a Voicing Cue for Syllable-Initial Stop Consonants. *Language and Speech* 22:173–89. https://doi.org/10.1177/002383097902200207.

Whalen, D. H., Arthur Abramson, Leigh Lisker, and Maria Mody. 1993. F0 gives voicing information even with unambiguous voice onset times. *The Journal of the Acoustical Society of America* 93:2152-2159. https://doi.org/10.1121/1.406678

Withgott, Mary Margaret. 1982. Segmental Evidence For Phonological Constituents. Proquest Dissertations Publishing.

Appendix A: Production Study Test Words

Condition	Environment description	Consonant	Word
1	At the beginning of a syllable with primary stress; word-initial.	k	capital
			caravan
			conduit
			consonant
		p	paradise
			positive
			powerful
		t	tapestry
			telescope
			tournament
2	At the beginning of a syllable with primary stress; non-initial.	k	decanter
			recapture
			recover
		p	appearance
			propulsion
			repeater
			supporter
		t	atomic
			determine
			interpret
3	At the beginning of a syllable with secondary stress; word-initial. **Three-syllable**	k	clarinet
			concierge
			condescend
			kangaroo
		p	personnel
			Portuguese
			preconceive
		t	tambourine
			tangerine
			Tennessee

Condition	Environment description	Consonant	Word
4	At the beginning of a syllable with secondary stress; word-initial. **Five-syllable**	k	cafeteria
			complimentary
			kilimanjaro
		p	pandemonium
			paradoxical
			positivity
			possibilities
		t	technicality
			terminology
			testimonial
5	At the beginning of a syllable with secondary stress; non-initial.	k	consequence
			honeycomb
			indicate
			leprechaun
			motorcade
		p	Davenport
			lollipop
		t	advertise
			attitude
			overtime
6	At the beginning of a word-initial stressless syllable.	k	cabana
			casino
			concussion
			connection
		p	pacific
			pajamas
			potato
		t	tacoma
			tomato
			torpedo

Condition	Environment description	Consonant	Word
7	At the beginning of a stressless syllable when immediately preceded by a stressless syllable and followed by a stressed one.	k	abracadabra
			delicatessen
			telekinesis
		p	anthropomorphic
			lollapalooza
			multiplication
			teleportation
		t	dramatization
			Mediterranean
			opportunistic
8	In coda position at **any** stress level.	k	acne
			actress
		p	hypnosis
			lapse
		t	atlas
9	At the beginning of a stressless syllable following a stressed one.	k	Indianapolis
			matriculation
			nicotine
			racquet
		p	aspen
			duplication
			rapid
		t	entertain
			nautical
			rhetoric
10	As the possible second member of an initial onset to a stressed syllable.	k	skeleton
			skydiver
		p	spatula
			spectator
		t	stadium

Condition	Environment description	Consonant	Word
11	As the possible second member of an initial onset to a stressless syllable.	k	schematic
			skedaddle
		p	spaghetti
			sporadic
		t	statistics
12	As the possible second member of a non-initial onset to a stressed syllable.	k	discovery
			escape
		p	exposure
			inspire
		t	extinguish
13	As the possible second member of a non-initial onset to a stressless syllable.	k	casket
			outskirt
		p	exposition
			inspiration
		t	institute
14	At the beginning of the second syllable of a stress lapse at the end of a word.	k	America
			Antarctica
			basilica
			harmonica
		p	chemotherapy
			colonoscopy
			philanthropy
		t	electricity
			personality
			university

Appendix B: Production Study Question and Answer Frames

1. Which [WORD] did her sister unpack? Her sister unpacked the [WORD] that was badly damaged.
2. Which [WORD] did his friend stop? His friend stopped the [WORD] that rolled past him.
3. Which [WORD] did her sister polish? Her sister polished the [WORD] that was particularly dirty.
4. Which [WORD] did her sister unpack? Her sister unpacked the [WORD] that was badly damaged.
5. Which [WORD] did your sister make? My sister made the [WORD] that turned different colors.
6. Which [WORD] did her brother notice? Her brother noticed the [WORD] that smelled funny.
7. Which [WORD] did the boy examine? The boy examined the [WORD] that glowed in the dark.
8. Which [WORD] did your friend freeze? My friend froze the [WORD] that broke into pieces.
9. Which [WORD] did your friend watch? My friend watched the [WORD] that caught fire.
10. Which [WORD] did our dog step on? Our dog stepped on the [WORD] that faded over time.
11. Which [WORD] did the girl hug? The girl hugged the [WORD] that squeaked loudly.
12. Which [WORD] did her brother read? Her brother read the [WORD] that hung overhead.
13. Which [WORD] did your sister fix? My sister fixed the [WORD] that spun in a circle.
14. Which [WORD] did her friend wear? Her friend wore the [WORD] that had red stripes.
15. Which [WORD] did his friend discuss? His friend discussed the [WORD] that was in his hand.
16. Which [WORD] did your brother catch? My brother caught the [WORD] that flew through the air.
17. Which [WORD] did his cat scratch? His cat scratched the [WORD] that looked threatening.
18. Which [WORD] did her teacher observe? Her teacher observed the [WORD] that lay on the ground.
19. Which [WORD] did your family pet sleep on? My family pet slept on the [WORD] that radiated heat.
20. Which [WORD] did your brother kick? My brother kicked the [WORD] that landed at his feet.
21. Which [WORD] did her brother clean? Her brother cleaned the [WORD] that oozed green slime.
22. Which [WORD] did her doctor suggest? Her doctor suggested the [WORD] that came in a box.
23. Which [WORD] did your sister see? My sister saw the [WORD] that needed painting.

Appendix C: Categorical Perception Study Alternative Spellings

Condition	Environment description	Voiced equivalent	Conventional spelling
1	At the beginning of a syllable with primary stress; word-initial.	gonduit	conduit
		gapital	capital
		gonsonant	consonant
		garavan	caravan
		baradise	paradise
		bowerful	powerful
		bositive	positive
2	At the beginning of a syllable with primary stress; non-initial.	regover	recover
		deganter	decanter
		recagture[4]	recapture
		abbearance	appearance
		probulsion	propulsion
		subborter	supporter
		rebeater	repeater
3	At the beginning of a syllable with secondary stress; word-initial. **Three-syllable**	gangaroo	kangaroo
		goncierge	concierge
		glarinet	clarinet
		gondescend	condescend
		bortuguese	portuguese
		bersonnel	personnel
		breconceive	preconceive
4	At the beginning of a syllable with secondary stress; word-initial. **Five-syllable**	gilimanjaro	kilimanjaro
		gafeteria	cafeteria
		gomplimentary	complimentary
		baradoxical	paradoxical
		bossibilities	possibilities
		bandemonium	pandemonium
		bositivity	positivity
5	At the beginning of a syllable with secondary stress; non-initial.	motorgade	motorcade
		lepreghaun	leprechaun
		honeygomb	honeycomb

[4] The voiced equivalent for this word was misspelled, and should have been “recabture.”

		indigate	indicate
		davenbort	davenport
		lollibop	lollipop
6	At the beginning of a word-initial stressless syllable.	gabana	cabana
		gasino	casino
		goncussion	concussion
		gonnection	connection
		bajamas	pajamas
		bacific	pacific
7	At the beginning of a stressless syllable when immediately preceded by a stressless syllable and followed by a stressed one.	abragadabra	abracadabra
		teleginesis	telekinesis
		deligatessen	delicatessen
		lollabalooza	lollapalooza
		telebortation	teleportation
		anthrobomorphic	anthropomorphic
9	At the beginning of a stressless syllable following a stressed one.	raguet	racquet
		matrigulation	matriculation
		nigotine	nicotine
		dublication	duplication
		indianabolis	indianapolis
		asben	aspen
		rabid	rapid
14	At the beginning of the second syllable of a stress lapse at the end of a word.	basiliga	basilica
		ameriga	america
		antarctiga	antarctica
		harmoniga	harmonica
		colonoscoby	colonoscopy
		philanthroby	philanthropy
		chemotheraby	chemotherapy