Investigating the phonological representation of Canadian Raising

Maggie Baird
University of Massachusetts, Amherst
mbaird@umass.edu

Abstract

The phonological process of Canadian Raising, involving the vowels [aɪ] and [ʌɪ] has been the subject of much theoretical analysis due to its apparent opacity and implication for phonological theory. This opacity assumes that the vowels are in an allophonic relationship, with surface form [ʌɪ] being derived from allophone [aɪ]. Two experiments were conducted to test whether this assumption is correct or whether the vowels may instead contrast underlyingly. In both experiments, I tested if [ʌɪ] leads to stronger prediction about upcoming sounds than [aɪ] does, which would be predicted under the allophonic representation. In experiment 1, listeners heard gated stimuli and responded with what they believed to be the final consonant. In experiment 2, listeners heard cross-spliced stimuli and responded with what they heard as the final sound. Both experiments show [ʌɪ] makes a stronger prediction - participants were most accurate at responding to the [ʌɪ] stimuli in the gating task, and were slowed down the most by mismatched cross-spliced stimuli with [ʌɪ] + voiced stops as opposed to [aɪ] + voiceless stops in the second task. These results suggest that indeed [aɪ] is a phoneme and [ʌɪ] is a derived allophone.
Keywords: Canadian Raising, phonological representation, allophony, gating, cross-splicing

1

The relationship between the diphthongs [aɪ] and [ʌɪ] in Canadian English has been the subject of much phonological analysis. Traditional analyses posit that /aɪ/ is the underlying phoneme which raises to the allophone [ʌɪ] under certain phonological conditions. The process interacts opaquely with tapping in North American English. An alternate account of

the data is that both diphthongs /aɪ/ and /ʌɪ/ contrast underlyingly, which would eliminate the posited opacity in this dialect. This paper presents two experiments - a gating study 3 and a cross-splicing study 4 that empirically test the phonological representations of these two vowels. The results of these experiments showed that the phone [ʌɪ] makes a stronger prediction about the voicing of the next consonant than [aɪ], which is predicted under the traditional view of a phonemic /aɪ/ and allophonic [ʌɪ].

Canadian Raising refers to the distribution of the phones [aɪ] and [ʌɪ] in certain dialects of English. First described by Joos (1942), Canadian Raising is the process where the diphthongs [aɪ] and less frequently [aʊ] raise to [ʌɪ] and [ʌʊ], respectively, when followed by voiceless obstruents. The process has been well-studied because of its opaque interaction with flapping in the North American English, where [t, d] become [ɾ] in certain stress environments. In Canadian Raising dialects, we see patterns as in 1:

(1) a. 'write' [rʌɪt], 'writer' [rʌɪɾɚ]
b. 'ride' [raɪd], 'rider' [raɪɾɚ]

On the surface, there is complete contrast between these two vowels. 'rider' and 'writer' are a minimal pair, with both vowels [aɪ] and [ʌɪ] appearing in identical contexts. The traditional explanation of these data is treating raising and tapping as two processes in an opaque ordering relationship. In these analyses, the phoneme /aɪ/ raises to [ʌɪ] preceding voiceless obstruents and on the surface there is overapplication of the raising process before taps derived from [t]. A rule-ordering account (as in Harris (1951)) is presented in Table 1.

Table 1: Rule Ordering for Canadian Raising Opacity

	'rider'	'writer'
	/raɪd + ɚ/	/raɪt + ɚ/
Raising	–	rʌɪtɚ
Flapping	raɪɾɚ	rʌɪɾɚ
	[raɪɾɚ]	[rʌɪɾɚ]

Under a rule-ordering analysis, the raising process occurs before the flapping process and we have a case of counterbleeding opacity. If the rules were in the opposite order, flapping would have bled raising in 'writer' and we would not see any raised diphthongs on the surface. The environment that triggered the raising is not present on the surface.

Opacity cannot be represented in simple parallel grammars because opacity requires a serial derivation. In a standard Optimality Theory (OT; Prince and Smolensky (1993/2004))

representation, the surface form of 'writer' [rʌɪɾɚ] will never be optimal. It violates the phonology of the rest of the language where the raised vowel does not precede voiced obstruents.

Because of the difficulty of representing opacity in certain phonological theories, some analyses of Canadian Raising utilize a different representation than the one underlying phoneme /aɪ/ (Mielke, Armstrong, & Hume, 2003; Pater, 2014). These accounts posit that the two vowels contrast, either throughout the entire language, or positionally, in order to derive the surface patterns. There has been significant back-and-forth about this issue in the literature (Bermúdez-Otero, 2004; Hall, 2005; Idsardi, 2006; Nazarov, 2019). Throughout this paper, I will refer to the traditional analysis of phonemic /aɪ/ and derived allophone [ʌɪ] as the **abstract** hypothesis, and the analysis with underlying contrast between /aɪ/ and /ʌɪ/ as the **contrast** hypothesis.

Previous perception and production experiments are inconclusive about the status of these vowels, with some results suggesting a "quasi-phonemic" or "marginal" contrast among these two vowels, while other results suggest an abstract representation. Marginal contrast does not have a strict definition, but refers to sounds which have some characteristics of contrast and others of non-distinctness. This term doesn't make a specific theoretical claim about the representation of these phones, but rather describes the pattern of behavior shown by speakers and listeners.

Hualde, Luchkina, and Eager (2017) concluded that the vowels in Canadian Raising have a "quasi-phonemic status". In a production study in Chicago, they found a difference in the realization of the diphthongs both in terms of length and in terms of F1. In a follow-up perception study, were less accurate at hearing the difference between Canadian Raising pairs "rider/writer" than voicing pairs such as "bright/bride". They were more accurate in distinguishing the Canadian Raising pairs than pairs such as "medal/metal" which have only an underlying contrast and no surface contrast, however. Hualde et al. (2017) note that participants clearly have an awareness of the difference in the diphthongs, but conclude the contrast is 'marginal' due to the lower accuracy and slower RT as compared to the voicing pairs.

Graham (2019) also concluded that these vowels have a marginal contrast in a novel paradigm she introduces of phonetic accommodation. (Graham, 2019) compares Predicted Raised (PR) words such as 'whiter' with Predicted Unraised (PU) words such as 'rider'. Participants read aloud a poem where two lines are strongly biased towards rhyming. The two lines ended in a PU word then PR, or vice-versa. Raising speakers should not rhyme these pairs, but the context biased towards a rhyming scheme. (Graham, 2019) tested the difference in duration and F1 and F2 based on vowel type and poetic placement. Participants

adjusted duration and F2 to match between the rhyming positions, but maintained the contrast between the two vowels in F1, the primary dimension of raising. Participant's unwillingness to change the F1 of the diphthongs suggests a contrast. Further investigation into individual participants did find some accommodation of F1. Graham concludes that certain dialects do have a marginal contrast of these two vowels, although the overall evidence is still murky.

Farris-Trimble and Tessier (2019), on the other hand, find evidence for an abstract representation using an eye-tracking paradigm. The logic of their study relies on the assumption of phonological inference (Gaskell & Marslen-Wilson 1996). Phonological inference was originally proposed in contrast with underspecification (CITE) and proposes that listeners accept allophones when they are in a licensed environment, as opposed to in any environment. There is mixed evidence (Gaskell & Marslen-Wilson 1998) if this is a completely prelexical phonological processor or if the inference mechanism interacts with the lexicon. In Gaskell and Marslen-Wilson (1998), they lay out two predictions about timecourse and phonological inference. First, an allophone, especially in assimilation contexts, can strongly predict a following segment because it is only licensed before that segment. Their paper presents phoneme monitoring experiments for cross-word coronal place assimilation, so this hypothesis predicts faster monitoring of the triggering segment (e.g. the /b/ in a /p b/ sequence). The other hypothesis is that the "undoing" of the phonological change takes time. This would predict a slowdown for phoneme monitoring as the listener is delayed undoing process (e.g. turning the [p] into a [t] in a /p b/ sequence). Farris-Trimble and Tessier (2019) take the second timecourse prediction as their hypothesis. In essence, the more transformations one has to undo, the slower recognition will be. Under a representation of these vowels where they contrast underlyingly, there is no phonological change being undone and recognition of either phone will be equally fast. Listeners have direct access to surface forms and do not have the intervening level of phonological inference. With direct access, listeners should be fast at recognizing surface forms which are heard often.

In Farris-Trimble and Tessier (2019)'s experiment, they created four sets of words in quadruplets: unraised ('bible'), raised ('biking'), opaque ('biting') and unrelated ('stapler').The recordings of the words were edited to equalize vowel length in order remove the cue of vowel length to voicing. In terms of phonological inference, the key difference is that the unraised have 0 phonological processes, the raised have 1 phonological process (raising) and the opaque have two phonological processes (raising and tapping). With an abstract representation and phonological inference, lexical recognition is predicted to be slower for each phonological process. In the experiment, participants were familiarized

with the association between the words and target images. The 4 images for a quadruplet were presented and participants heard one of the words and were tasked with clicking on it, although eye-tracking data was the primary measure. The first fixation time was fastest for the unraised condition followed by raised and then opaque. In their analysis, each of the pairwise interactions (unraised vs raised/opaque, and raised vs. opaque) were significant. They also looked at the target midpoint - the point at which 1/2 of all fixations were to the target item. In this case, the fastest was filler, then unraised, then raised, and finally opaque. These results suggested that a greater number of phonological processes leads to slowdown in lexical recognition. Based on these experiments, Farris-Trimble and Tessier found support for the abstract/opaque analysis of Canadian Raising, where the raised diphthong is an allophone of the unraised vowel and is derived through opacity in the phonological grammar. The results do not support a contrast/direct access model where the number of phonological processes should not impact lexical recognition times.

2 Experimental Design

The linking hypothesis in Farris-Trimble and Tessier (2019) is that undoing phonological transformations takes time, and therefore the more phonological changes in a word, the slower lexical recognition is. This is one dimension of phonological inference as it was proposed in Gaskell and Marslen-Wilson (1998), but there is another idea put forth in that paper. Gaskell and Marslen-Wilson (1998) propose that viable phonological changes could lead to predictions about the context that causes that change. In essence, when a listener hears an allophone, they are able to predict the following sound that triggered that allophony. This theory proposes a speed-up effect in certain tasks involving allophony, as opposed to the inhibitory effect of undoing phonology. This idea was tested in Gaskell and Marslen-Wilson (1998) with a phoneme-monitoring study for the triggering consonant for coronal-place assimilation. The results were inconclusive, but follow-up work by Gow (2001) found that participants have a faster reaction time to a triggering consonant when preceded by an assimilated consonant than an unassimilated one. Both effects proposed by Gaskell and Marslen-Wilson (1998) rely on an abstract representation as opposed to direct access. The experiments in the current paper test the proposal that allophony leads to prediction of the upcoming triggering environment and leads to a speed-up in certain tasks. The experiments in this paper serve both to test the representations of the vowels in Canadian Raising and also to validate the hypothesis of prediction-based processing of allophony as a component of phonological inference.

The experiments in this study are versions of a gating task (Smits, Warner, McQueen,

& Cutler, 2003) and a subcategorical mismatch task (McQueen, Norris, & Cutler, 1999). In the gating task, participants heard words (e.g. *write, ride*) and nonwords (e.g. *gite, kide*) with either a raised or unraised diphthong followed by a word-final voiced or voiceless consonant and were tasked with identifying the last consonant. The stimulus was cut off at different intervals throughout the nucleus. The primary purpose of the gating task was to confirm that speakers with Canadian Raising are able to differentiate these diphthongs and use height information to make predictions about the upcoming final consonant. Under the abstract hypothesis, we expect accuracy to increase on further gates and also expect higher accuracy earlier for the raised diphthong/voiceless consonant, as the allophone makes a stronger prediction about the upcoming segment. Under a contrast analysis, predictability of the final consonant should not be categorically different between the diphthongs, but rather be based on some distributional properties of the language and sounds.

The second experiment was a cross-splicing experiment. In this experiment, participants heard full words or non-words (same words as exp 1), but the final consonant was spliced from a different token. The final consonant would either be the same as the original pronunciation or have the opposite voicing. Under either hypothesis, the cross-spliced items should have longer response times than identity-spliced items as they are phonologically incongruous. Under the abstract hypothesis, within the cross-spliced items, raised vowels would have longer response times than the unraised vowels. The raised vowel is the allophone and therefore predicts an upcoming voiceless consonant. When that prediction is not met, slow response times are expected. The unraised vowel, the phoneme, does not make the same prediction about the voicing of the upcoming consonant. Under the contrast analysis, I did not expect a difference in response time between the two mismatch cases as both are equally ungrammatical.

Depending on the formal theory attached to the contrast hypothesis, the allophone should still be predictive of the upcoming consonant because of distributional information. This idea is explored further in section 5, but the crucial distinction is that the abstract hypothesis predicts an asymmetry in predictiveness of each vowel where contrast does not. Under a contrast representation, each vowel can only be followed by certain consonants and each vowel therefore predicts upcoming material. Under an abstract representation, the prediction is that the allophonic raised vowel is making a stronger prediction than the phonemic unraised vowel. This prediction comes from two facets of the theory. First, as described above, the voiceless obstruent is the triggering environment for the allophony so the allophone can only be present when a voiceless obstruent is immediately following. Second, the unraised vowel has phonemic status and is therefore the unit of representation in the lexicon. The lexical for 'write', for example, contains the phoneme [aɪ]. Therefore,

with abstract lexical representations and distinct surface forms, the presence of [aɪ] in a word such as 'write' is less anomalous than the presence of [ʌɪ] in a word such as 'ride' because 'write' does contain /aɪ/ at a phonemic level.

In order to confirm the validity of the linking hypothesis utilized in these experiments, I also tested a set of control items involving nasal deletion. In English, vowel-nasal sequences undergo nasal deletion preceding voiceless consonants and are realized as a nasalized vowel. Before voiced consonants, the nasal surfaces faithfully as a consonant. Cohn (1993) compared these two realizations and showed that anticipatory nasalization of a vowel preceding a nasal consonant is a gradient phonetic realization of nasalization. The nasalization shows interpolation over the span and not as a categorical feature realization. Nasal Deletion, on the other hand, she showed to be phonological, with nasal airflow having a plateau across the vowel and not interpolation. This suggests that the vowel has a [+nasal] feature. Krämer (2019) in a review of nasalization literature also finds that nasalization in English is phonological rather than phonetic. We can be quite confident that vowel nasalization in these environments in English is not contrastive and is phonological derived. The nasalization and deletion process is also conditioned by a following voiceless obstruent, and therefore serves as a direct comparison to Canadian Raising. Thus, the nasal deletion items serve to validate the experiments as I am interpreting the tasks in a novel way. In the gating experiment, the Nasal stimuli should show higher accuracy at earlier gates for the nasalized vowel/voiceless stimuli as the vocalic allophony predicts the upcoming consonant. In the cross-splicing experiment, there should be a longer reaction time in the mismatch [$\tilde{V}$] + [+voice] items as compared to the mismatch [Vn] + [-voice] items as the stronger prediction comes from the allophonic nasalized vowel will lead to a significant slowdown when it is not met. If these patterns are found, the predicted behavior of allophones in these tasks is confirmed and analysis of the Canadian Raising as hypothesized can occur.

In addition to utilizing a distinct linking hypothesis, the experiments presented here also address some of the limitations of Farris-Trimble and Tessier (2019). First, the stimulus set was quite small because of how constrained they had to be in designing the quadruplets and the nature of the English lexicon. The stimulus set in the used in these experiments is much larger, with 58 total Canadian Raising words tested. Second, there was a confound with morphological complexity. Nearly all (11/12) of the opaque words were morphologically complex (e.g. biting, sighting), whereas significantly fewer of the unraised (6/12) and raised words (6/12) were monomorphemic (e.g. Nike, license). The morphological complexity could also add to the word recognition time and is confounded with the number of phonological processes. The current study uses exclusively monosyllabic words and

overwhelmingly monomorphemic words, the only exceptions being a few inflectional past tense and plural suffixes. Finally, the imageability of the words was very variable as well as their frequencies. Although the authors did familiarize the participants with the word-image connections, there were large differences in the imageability. For example, in the quadruple I have used as an example, 'bible' and 'biking' are quite clear, even though 'biking' is a verb and 'bible' is a noun. The image for 'biting', however, was a child eating a watermelon, where the watermelon was very salient but the target was more abstract. It is possible that the opaque images were overall more difficult to recognize, adding to the recognition time. This study uses exclusively audio stimuli and avoids frequency discrepancies as much as possible.

3 Experiment 1

Experiment 1 was a gating task, where participants heard a partial or complete stimulus followed by a square wave and were tasked with responding what they believed the final sound was. Stimuli creation was based on Smits et al. (2003), which investigated the recognition of diphones in Dutch.

3.1 Methods

3.1.1 Stimuli

The stimuli for both experiments were drawn from the same set of monosyllabic words (see Appendix A). The words all ended in the consonants [t/d, p/b, f/v, s/z], chosen for the voicing contrast and lexical availability. The stimuli for the gating study comprised 48 total items - 16 Canadian Raising, 16 Nasal, and 16 fillers.

Within the Canadian Raising stimuli, there were 6 items where both the voiced and voiceless version of the final consonant made a real English word (e.g. *hide/height*), 5 where only the voiced consonant made a word (e.g. *guide/gite*) and 5 where only the voiceless consonant made a word (e.g. *kite/kide*). Throughout the paper, these lexical statuses will be referred to as 'both', 'only voiced' and 'only voiceless'. The Canadian Raising set posed the biggest challenge because of the limits of the English lexicon. For the pairs, there were a total of 26 candidates in the English lexicon, and I selected those whose frequency ratios (taken from SUBTLEX) were the least extreme. The items where only one voicing made a word were selected to provide the most balanced selection of final consonants, although equality is not possible in the lexicon. This stimulus set was larger than that of Farris-Trimble and Tessier (2019) to hopefully garner a robust view of the behavior of this

process. In addition, morphological complexity was avoided wherever possible, with only one morphologically complex word included (*lied*).

The Nasal stimuli were chosen in a very similar way. All of these words ended in [nt/nd], so the final consonant balancing is not a concern. There were 6 items where both [nt] and [nd] made a word (e.g. *grant/grand*). Like the Canadian Raising cases, these were chosen based on their frequency ratios. There were 5 of each of the 'single' cases, where only [t] made a word (e.g. *slant/slanned*) and where only [d] made a word (e.g. *blend/blent*). These were chosen to be of non-extreme frequency and also not past tense. The use of the morphologically complex past tense could not be completely avoided and was present in 3 items (*joined, planned, tinned*).

The fillers were monosyllabic words ending in [p, b, f, v, s, z]. There were not fillers ending in [t, d] because those consonants were overrepresented in the data due all of the nasal stimuli ending in [t, d]. There were 16 total fillers - 6 [p, b]-, 5 [s, z]- and 5 [f, v]-final words. The fillers had a mix of wordedness, with instances of each voicing making a word as well as both.

All of the words and nonwords were recorded by a native Canadian English speaker in a soundbooth with a SHURE SM35 Performance Headset microphone. The stimuli were produced naturally but with clearly enunciated voicing on the final consonant. Formant trajectory analysis showed a significant different between raised and unraised vowels in the natural speech and additional recordings with exaggerated raising were not used. Unlike Farris-Trimble and Tessier (2019), vowel length was not normalized. This was done for two reasons. First, to keep the stimuli as naturalistic as possible. Second, as the gating was done by percentages, the vowel length differences at earlier gates should not be hugely informative as different vowels and tokens had different lengths to begin with.

Each item had a total of 8 versions - 2 final consonants x 4 gates. All gates were determined based on percentiles of the nucleus. For the Canadian raising items and the fillers, the percentile was calculated just from the vowel. For the nasal cases, things were a little bit different. For words ending in [nt], there was often no nasal consonant present at all in the acoustics, and the nasalization was fully realized on the vowel. For words ending in [nd], there was a full vowel and nasal consonant before the final consonant. For making the gates, I treated all the material beginning with the vowel and ending before the final consonant as the nucleus. So, for the [nt] cases, this was often just a vowel, but for the [nd] cases, this was a vowel and a nasal. Thus, the third gate always included all material up to but not including the final target consonant. All stimuli were created and edited using Praat (Boersma & Weenik 2001). Gated stimuli were created by calculating nucleus length and cutting off the stimuli at specific points.The first gate was 1/3 of the nucleus,

the second gate was 2/3 of the nucleus, the 3rd gate was the whole nucleus but not the final consonant, and the 4th gate was the entire stimulus. After the cutoff point, each stimuli was followed by a square wave with an F0 of 500Hz. The length of the beep was such that the total length of each stimulus is 1.0 seconds. There was a beep following the 4th gate as well, after the final consonant. The purpose of the beep was both to standardize the length of the stimuli and also to provide acoustic material after the stimulus to avoid any perceptual effects that may arise from the cutoff. After the creation of the gated stimuli, all the stimuli were scaled to a peak intensity of 0.97. The addition of the square wave as well as a 5ms ramp-down of the stimulus followed Smits et al. (2003) procedure in order to prevent truncation of the stimuli which may have affected responses. Samples of the stimuli can be viewed in **??**.

Each participant heard 4/8 versions of each item. This was to prevent participants from realizing they were hearing pairs of words and also to keep the experiment a reasonable length.Each participant saw one case of each gate: 2 from the voiced consonant, and 2 from the voiceless consonant. For example, for the item "lied/light", a single participant would see either Gate 1 and 3 of lied and Gate 2 and 4 of light or Gate 1 and 3 of lied and Gate 2 and 4 of light. This particular division pattern was to maximize distinctiveness in the stimuli heard by individual participants. Participants didn't see both voicings at the same gate to mitigate any kind of strategy or explicit awareness of the contrasts present in the study. The study would have also become quite long and tiring with the full stimulus set.

3.1.2 Procedure

The experiment was run online using PCIbex software (Zehr and Scwarz 2018). Participants first recorded themselves reading a paragraph (see Appendix B) which was used to confirm they spoke a raising dialect. They were then instructed that they would be hearing a real or fake word and have to respond what the last sound was. They were told to give their best guess if the last sound is difficult to hear. They were instructed to wear headphones.

Items were presented in a random order and there were no practice trials. At the beginning of each trial, a fixation cross would appear for 500ms. Next, the two consonant response options for that trial would appear as letters on the screen under the test "press <- for" and "press -> for" for 1000ms. The voiced consonant always corresponded to the left arrow and the voiceless consonant corresponded to the right arrow. This consistency was to keep relative consistency for the participants and not confuse them. Because the arrow keys are both on the same hand, slowdown from the non-dominant hand was not a concern. Next, the audio file played and participants responded with a keyboard press to

what they believed the last sound to be.

3.1.3 Participants

Participants were recruited from Prolific and were paid 2.50 USD for their participation. All participants were self-reported monolingual English speakers and were born in and currently live in Canada. The first experiment had a total of 30 participants. The recordings of the initial paragraph were listened to by a trained Canadian linguist and all clearly had raising as a part of their dialect.

3.2 Results

For data analysis purposes, responses between 150ms after offset and 5000ms after offset were considered. These ranges were chosen generously to exclude errors such as misclicks and trials where the participant looked away or wasn't paying attention. This excludes 4.3% of the total responses. Participants were not told specifically to respond as quickly as possible, and as such only unreasonable response times were excluded. The total reported range of response times varied from 80ms to 96,176ms. Fillers are excluded from analysis in both experiments and only served to distract participants from ascertaining a pattern in the experiment.

3.2.1 Accuracy

First, I will consider accuracy. Figure 1 shows a simple representation of accuracy by condition and gate with standard error bars.

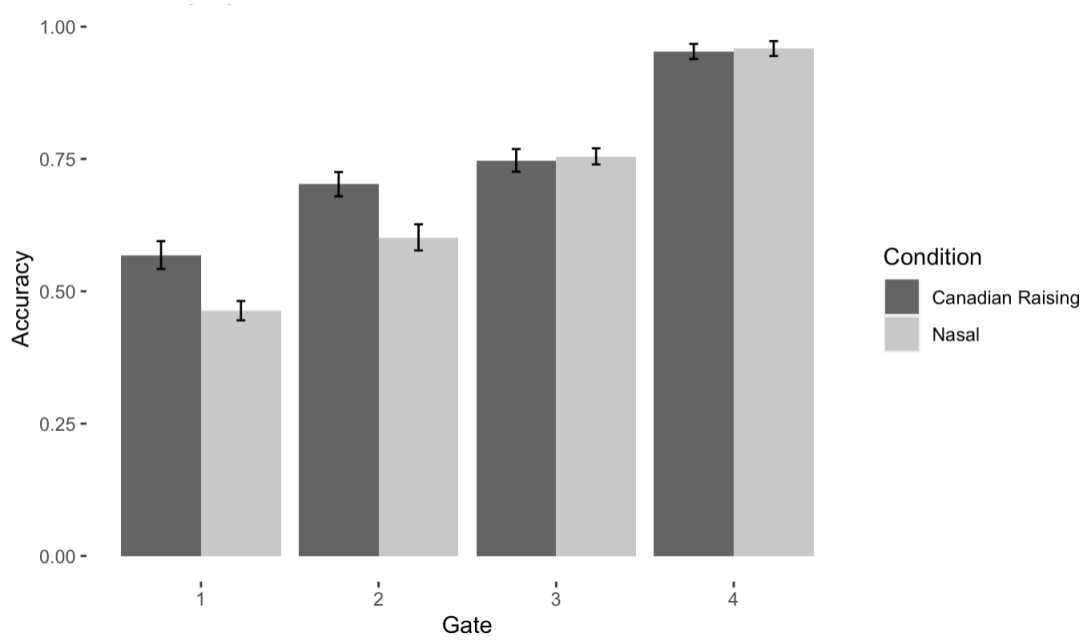

Figure 1: Exp 1 accuracy at each gate for CR and Nasal stimuli

As expected, participants had higher accuracy with further gates. On the first two gates, accuracy was higher for the Canadian Raising stimuli than the nasal stimuli. For the third and fourth gates, there did not appear to be a difference between the two conditions. At the fourth gate, there was near ceiling accuracy, as participants were hearing the consonant that they are responding to.

Figure 2 shows accuracy by gate, condition, and voicing.

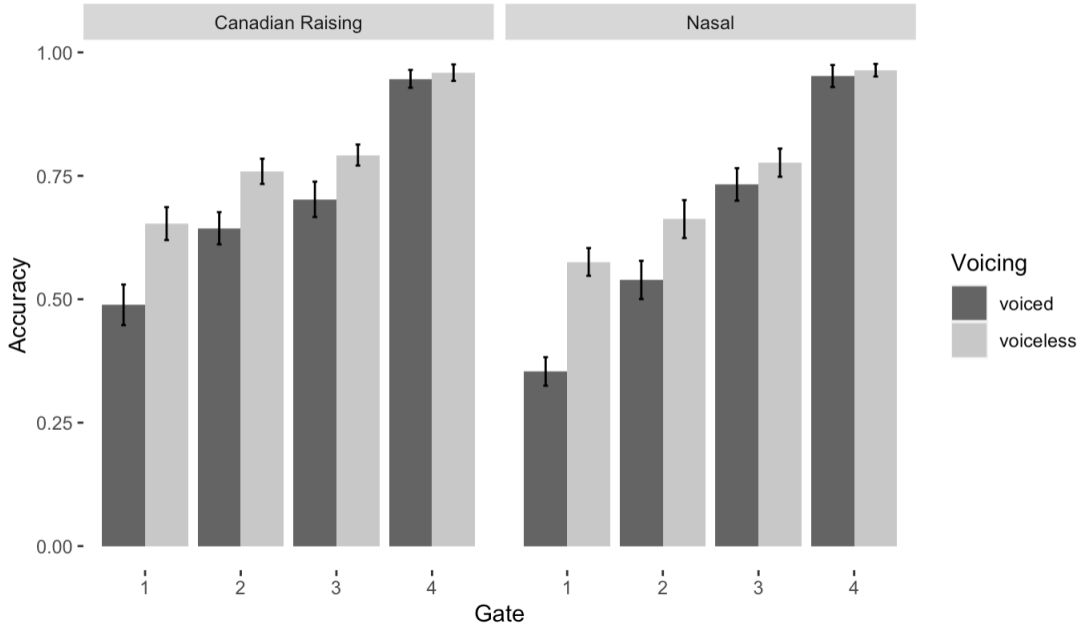

Figure 2: Exp 1 accuracy at each gate, separated by condition and voicing of the correct response

Figure 2 shows that voiceless consonants had a higher response accuracy than the voiced consonants in the Canadian Raising and nasal cases for the first 3 gates.

Figure 3 further divides accuracy into lexical status. For some pairs, both the voiced and voiceless consonant made a word (e.g. right/ride) - these are coded as 'Both'. Items where only the voiced consonant makes a word (e.g. guide/gite) are labelled as 'Only Voiced' and items where only the voiceless consonant makes a word (e.g. kite/kide) are labelled as 'Only Voiceless'. In Figure 3, stimuli are separated into their lexical status in the vertical panels. Voicing of the final consonant is shown at each gate with a light gray for the voiceless consonant and dark gray for the voiced consonant. Note that a voiced consonant in the only voiceless condition corresponds to a nonword (e.g. kide), and vice versa.

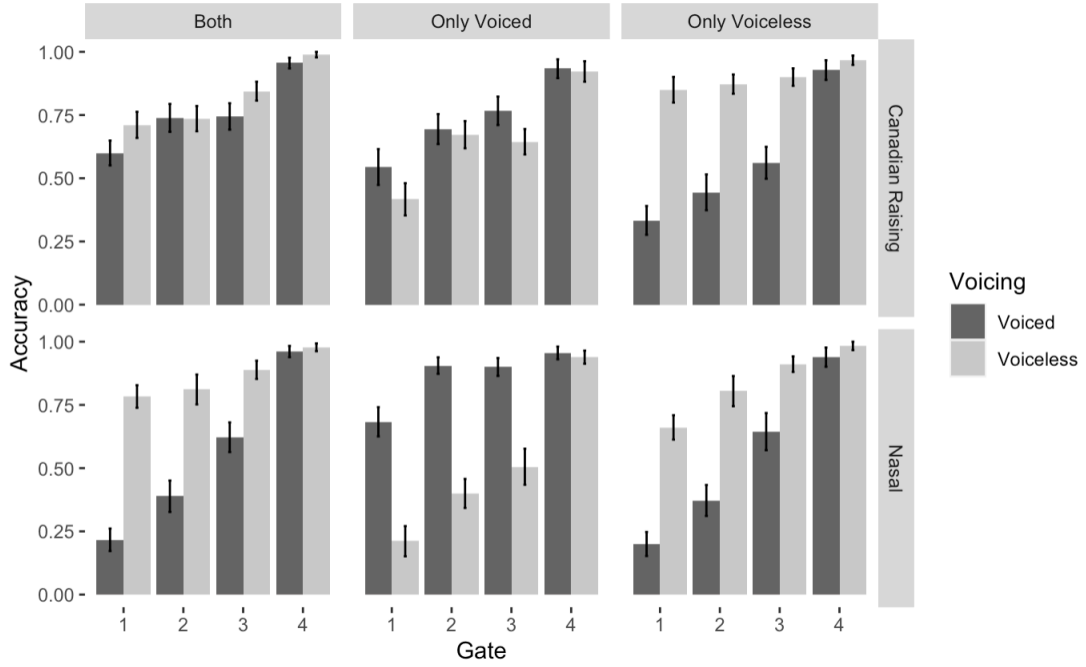

Figure 3: Exp 1 accuracy at each gate, separated by condition, voicing, and lexical status of the stimulus

For cases where both voicings made a word, the nasal stimuli showed a bias towards the voiceless response, especially in the first three gates. For the Canadian Raising stimuli, the effect was in the same direction, although the bias appears to be much weaker. For the cases where only the voiced consonant makes a word, both conditions saw a bias toward the voiced response in the first three gates, although the bias appears stronger for the nasals. For the cases where only the voiceless consonant makes a word, both conditions showed a strong bias toward the voiceless response at the first three gates.

3.2.2 Modelling

Results were analyzed with a mixed effects logistic regression model of the probability of a correct response using the glmer() function in R. Fixed effects included gate, voicing . Random effects included subject and item. Gates were coded as 1, 2, 3, 4. Voicing was contrast coded with voiced consonant coded as -.5 and voiceless consonants coded as .5. Lexical status was coded as follows: only voiceless = .5, only voiced = -.5, both = 0. The model also included an interaction between voicing and lexical status, based on qualitative patterns in the results. Models were run separately for each condition. Parameter estimates are shown in Tables 2 and 3.

Table 2: Mixed-Effects Model Parameter Estimates, Canadian Raising

	Estimate	Standard Error	Z value	p value
(Intercept)	-0.5115	0.1653	-3.094	< 0.002**
Gate	0.7436	0.0577	12.897	< 0.001***
Voicing	0.6903	0.1227	5.625	< 0.001***
Lexical	0.3884	0.2486	1.562	0.118
Voicing:Lexical	2.5810	0.3102	8.322	< 0.001 ***

Table 3: Mixed-Effects Model Parameter Estimates,Nasals

	Estimate	Standard Error	Z value	p value
(Intercept)	-1.3662	0.1513	-9.028	< 0.001***
Gate	1.0159	0.0618	16.437	< 0.001***
Voicing	0.7277	0.1245	5.846	< 0.001***
Lexical	0.2811	0.1930	1.456	0.145
Voicing:Lexical	4.058	0.3324	12.210	< 0.001***

As expected, there was a positive relationship between gate and accuracy, and also a positive relationship between voiceless consonants and accuracy. There was not a significant effect of lexical status, although there was an interaction between voicing and lexical status, with significantly higher accuracy for voiceless consonants in the only voiceless cases. The Canadian Raising and Nasal stimuli showed the same effects.

3.2.3 Response Time

Participants were not explicitly told to answer as quickly as possible. This is partially because of the difficulty of the task -the response consonants often changed from trial to trial, and the participants were also responding to the "sounds" that the letters represented, which may not have been a natural task. In particular, many /z/-final words end in <s> (e.g. spies). In addition, many words are orthographically vowel-final (e.g. bride), which may have been another difficulty for participants.

Response times were also recorded and investigated, and Figure 6 shows a summary of RTs by gate, condition, and voicing.

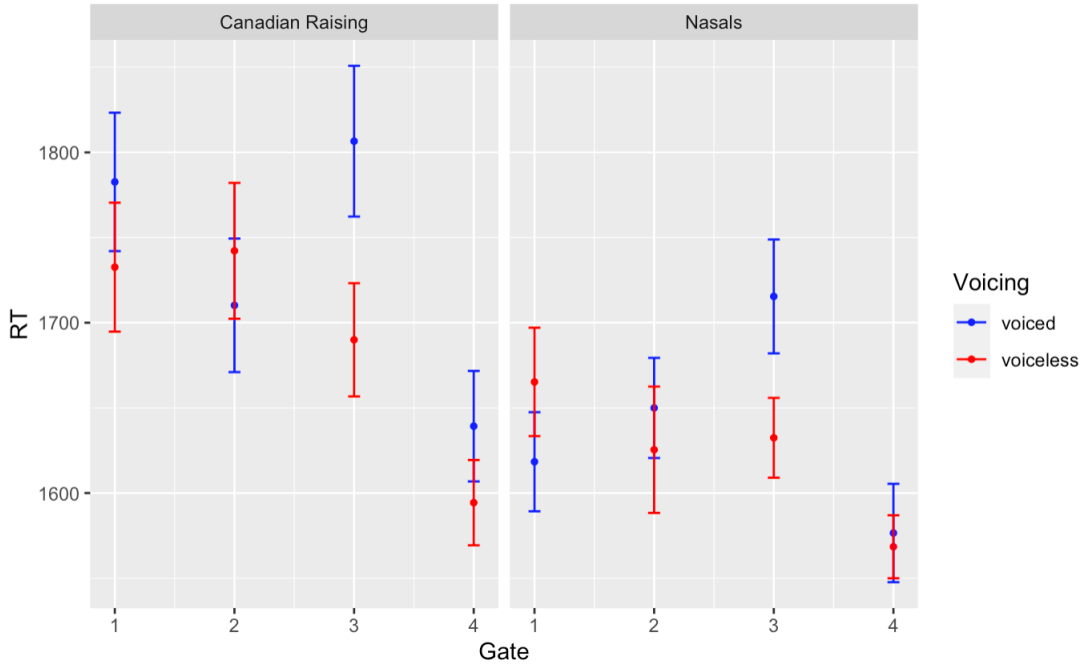

Figure 4: Exp 1 RT (ms) at each gate, conditions in separate panels and voicing of the final consonant shown by color

The only reliable pattern is that the 4th gate corresponded to the shortest response times. Recall that in the 4th gate, the participants are hearing the entire final consonant, so their decision should have been the easiest. It appears that the Canadian Raising stimuli led to slower response times overall. There is an interesting pattern where at the third gate, voiced stimuli were responded to slower than voiceless. At the third gate, participants were hearing the entire vowel. This qualitative result does align with the abstract hypothesis that the vowel is making a stronger prediction when it is an allophone, for the voiceless stimuli. RT did not track with accuracy, as voiceless responses had higher accuracy across all gates, but did not consistently show faster RTs. Response time was not the most informative measure in this study, and the lack of patterns in the first 3 gates is not of concern.

Keyboard presses in an online experiment are not as sensitive as other RT-based experimental setups in the lab. The short response times on the 4th gate and general trend of decreasing RT with gate did show a reasonable level of sensitivity in the experimental setup that is necessary for experiment 2.

3.3 Discussion

In all conditions, participants had higher accuracy as the gates increased, which validates the task. For the nasal condition, accuracy was higher for the voiceless consonant at each gate, especially the early gates, confirming the hypothesis of stronger prediction from the allophone. In the first two gates, which represented 1/3 of the vowel and 2/3 of the vowel, respectively, the participants were more accurate at the Canadian Raising stimuli than the nasal stimuli. This is likely because the formants of the raised and unraised vowels were immediately apparent and were able to distinguish the phones. This behavior supports an abstract, allophonic account, where the height of the vowel influenced people's predictions about the upcoming consonant. The lower accuracy on the nasal stimuli at the earlier gates could have stemmed from the phonetic cues of the allophones being less salient. There would be nasalization on the vowel in both cases, although to a lesser degree for the voiced consonants. The more salient phonological difference was the presence or absence of the nasal consonant, which would be apparent by the third gate, where the accuracy on the nasals caught up to the diphthongs.

In both conditions, when only the voiceless consonant made a word, there was a strong bias towards picking the voiceless consonant. In the only voiced case, there was a preference towards selecting the voiced consonant in each condition, although the effect was much more stark in the nasal case than the Canadian Raising. There was still relatively high accuracy on the voiceless cases for the Canadian Raising when only the voiced consonant made a word. In addition, the accuracy on voiceless is higher in the 'both' condition of Nasals. It seems the phonology strongly predicts a voiceless consonant, even if there is no lexical support. Under the abstract account, the voiceless words contain the strongly predictive allophone, and therefore a bias towards picking the voiceless consonant is expected.

In the case where both consonants made a word, there was a difference between the Canadian Raising and the nasal conditions. In the Canadian Raising, the accuracies were similar for each consonant at each gate, but for the nasal condition, there was still a bias towards selecting the voiceless consonant. This pattern is puzzling and not necessarily expected. It is unlikely that this was a frequency effect, as the -nt final words had an average frequency of 26.12 and the -nd final words 23.50. There was a bigger difference

in the median frequency, with the -nt words having a median of 9.27 and the -nd 3.71. It is important, though, that of the 19 total pairs of "both" nasal stimuli, 10 had the -nt word being higher frequency, and 9 the -nd word.

The results from the mixed effects model confirmed the story I saw with the visualization of the data. Accuracy was higher for voiceless consonants overall. This suggests that the vowel preceding a voiceless consonant is giving a stronger prediction about the upcoming consonant, which is in line with the allophonic, opaque analysis of Canadian Raising.

The lexical status of the stimuli is expected to bias the responses, and there was a puzzling difference between the nasal and the Canadian Raising stimuli in that regard. Although the overall results of this gating experiment seemed to align with an allophonic, abstract account of these vowels, further confirmation is needed about the predictions people have upon hearing these vowels. The cross-splicing paradigm is designed explicitly to test listener's predictions of upcoming material based on allophonic status. The paradigm controls for lexical status and measures response time.

4 Experiment 2

This experiment took methods from McQueen et al. (1999). Their experiment looked at words and nonwords in English and Dutch, and the manipulation was in the place of articulation of the final consonant. The vowels in their study were not allophones, as in the present study, but did differ in their formant behavior as a function of the upcoming consonant. Over the course of several experiments, they showed that participants were sensitive to this fine-grained phonetic detail and showed higher accuracy on identity-spliced items as opposed to cross-spliced items. They also found an effect of the lexicon, with different behavior for words and nonwords.

4.1 Methods

4.1.1 Stimuli

The wordlist for the Canadian Raising experiment was a superset of the stimuli for the gating experiment. As such, all recording details were identical as described in Experiment 1. The total set of words is 177 (58 Canadian Raising, 58 Nasal and 61 fillers). There were 58 Canadian raising items - 19 with 'both' lexcial status, 19 of the 'only voiceless' items, and 20 'only voiced'. There are 19 nasal items where both [t, d] made a word, 19 of the 'only voiceless' items, and 20 'only voiced'.

For the fillers, there were 21 [f, v] items, 19 [p, b] and 21 [s, z]. In the [f, v] items,

there were 5 in which each consonant made a word (e.g. *leaf/leave*), 8 where only [f] made a word (e.g. *cliff/cliv*), and 8 where only [v] made a word (e.g. *dove/duff*). In the [s, z] fillers, there were 8 words where each consonant made a word (e.g. *bus/buzz*), 7 where only [z] made a word (e.g. *jazz/jass*), and 6 where only [s] made a word (e.g. *mass/mazz*). In the [p, b] fillers, there were 9 words where each consonant made a word (e.g. *robe/rope*), 4 where only [p] made a word (e.g. *beep/beeb*), and 6 where only [b] made a word (e.g. *snob/snop*).

The full set of these stimuli were used in the cross-splicing experiment. For each item, there were 4 versions. These were V1C1, V1C2, V2C2 and V2C1. That is, I recorded each item with both consonants, whether they were words or not (e.g. *hide* and *height*, *kite* and *kide*). To create the stimuli, these recordings were separated into two parts - up to the end of the vowel, and the final consonant. Every stimulus that participants heard had its consonant spliced on. For identity splices (V1C1 and V2C2), consonants were taken from the same vocalic context - for example the [t] from *height* was spliced onto the CV of *kite*. The mismatch splices (V1C2 and V2C1) were also spliced from the same vocalic context, but not the same item (e.g. 't' from *kite* spliced onto intial material from *hide*.

All audio editing was done in Praat (Boersma & Weenik, 2001). During the splicing, a cosine ramp-down of 5ms was added at the spliced point to avoid any acoustic anomalies. As splices were between two different files, which will have different acoustic characteristics, the intensity of each component was also scaled. I calculated for each consonant type (voiceless stop, voiced stop, [f], [v], [s], [z]) its average intensity based on a random sample. The initial material (up to and including the vowel) was scaled to an intensity of 0.97 and the consonant was scaled to the ratio of its consonant type's average intensity to 0.97. Samples of the audio can be viewed in Appendix **??**.

4.1.2 Procedure

Participants began the experiment by reading the same paragraph as in Experiment 1 (see Appendix B). Participants were told that they would hear a real or fake word and be asked to respond what the last sound was. They were told to be accurate as well as quick and keep their fingers on the buttons. They were told to wear headphones.

Items were presented in a random order and there were no practice trials. For each trial, a fixation cross first appeared for 500ms. Next, the candidate consonants appeared as letters under the text "press <- for" and "press -> for" for 1000 ms. As with experiment 1, the left arrow key corresponded to the voiced consonant. Next, the audio file played and participants responded to what the final consonant was with a button press.

4.1.3 Participants

Participants were recruited on Prolific and were paid 2.50 USD for their participation. Participants were all self-reported monolingual speakers of English born and raised in Canada. A trained linguist listened to their recordings and confirmed that they all have raising in their dialects. 36 people participated.

4.2 Results

Initial exploratory analysis of the data showed that determining a response time cutoff could impact the significance of the effects under analysis. Notably, within the Canadian Raising results, the interaction between prediction (vowel type) and match was not always significant when different cutoff points were chosen for response times. I also noticed that the longest response times disproportionately belong to the voiceless predicted (allophone) + mismatch cases, which were also the cases that the interaction is picking out and the place where I expected to see an effect under an abstract analysis. In an effort to fully analyze the data and be transparent, I elected to perform a cross-validation analysis which is discussed in 4.2.3. For presentation and analysis, the only data excluded are those with a response time outside of 4 standard deviations of the mean log response time. This excludes 8 total trials of 4,176 (0.1%). The excluded trials had unreasonably long response times of less than 60ms and greater than 9130ms. As in experiment 1, participants were not told to respond as quickly as possible so I want to include long response times, though not those were it seems that participants took a break or were away from their computer.

4.2.1 Accuracy

In this experiment, participants were responding to what they hear as the final consonant of the word. They were able to hear the consonant in every case, so accuracy is not the primary measure, but is still informative.

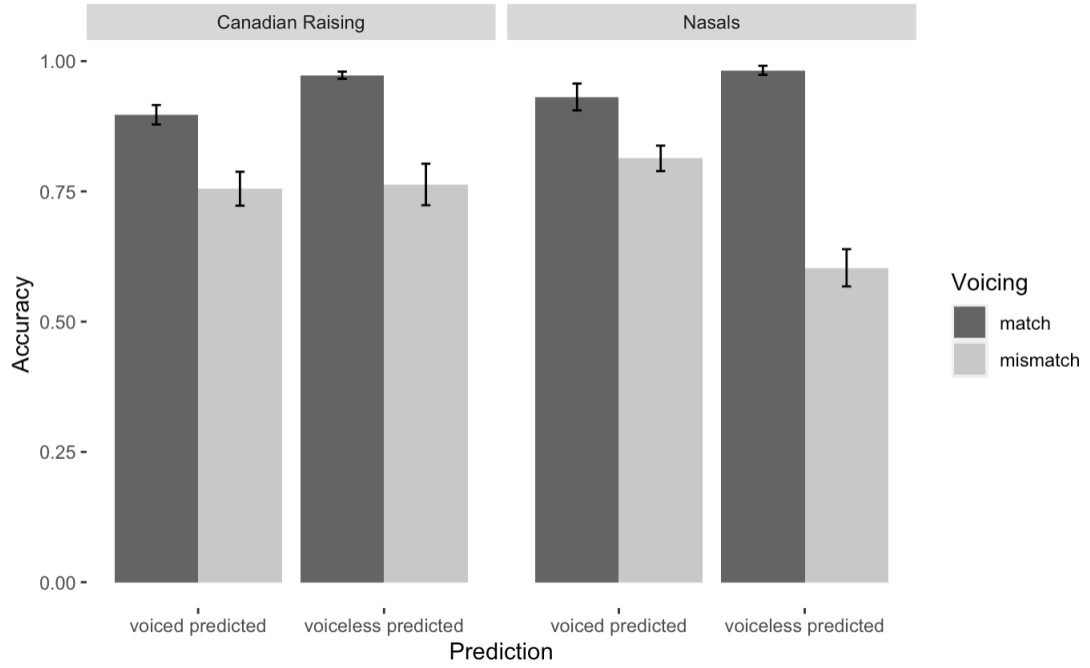

Figure 5: Exp 2 accuracy for each condition, separated by vowel and match, details in text

In Figure 5, the x axis represents the vowel. For Canadian Raising, "voiced predicted" is an unraised vowel and "voiceless predicted" is a raised vowel. For the nasals "voiced predicted" is a vowel + nasal sequence, and "voiceless predicted" is a nasalized vowel without a consonant. The color represents the actual voicing of the spliced consonant. For "voiced predicted", the dark grey matches are voiced consonants and for "voiceless predicted" the dark grey matches are voiceless consonants.

In Figure 5, there was highest accuracy when the prediction from the vowel matches the voicing of the consonant. In the nasal cases, there was lowest accuracy when the vowel predicts a voiceless consonant, but the consonant doesn't match. There wasn't the same effect with the Canadian Raising stimuli - the accuracy for both mismatch cases was roughly equivalent.

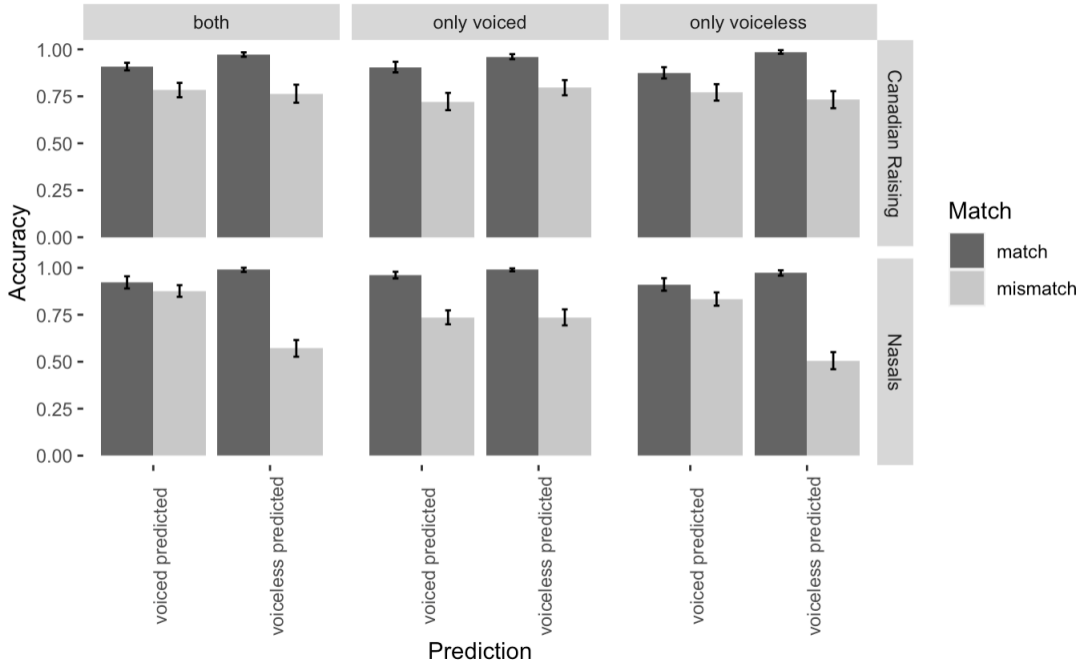

Figure 6: Exp 2 accuracy, separated horizontally by condition, vertically by lexical status of the stimulus. The vowel is characterized by 'prediction' as in other figures and consonant voicing shown by the color of 'match'.

Figure 6 shows that the lexical status had a moderate effect on accuracy, particularly for the nasal cases. When both consonants made a word and when only the voiceless consonant made a word, and participants heard a nasalized vowel without a nasal consonant, they respond with 't' even when they were hearing 'd'. There was not the same effect in the Canadian Raising cases.

4.2.2 Response Time

The primary dependent measure in this study was response time. Figure 7 shows response time by prediction and match.

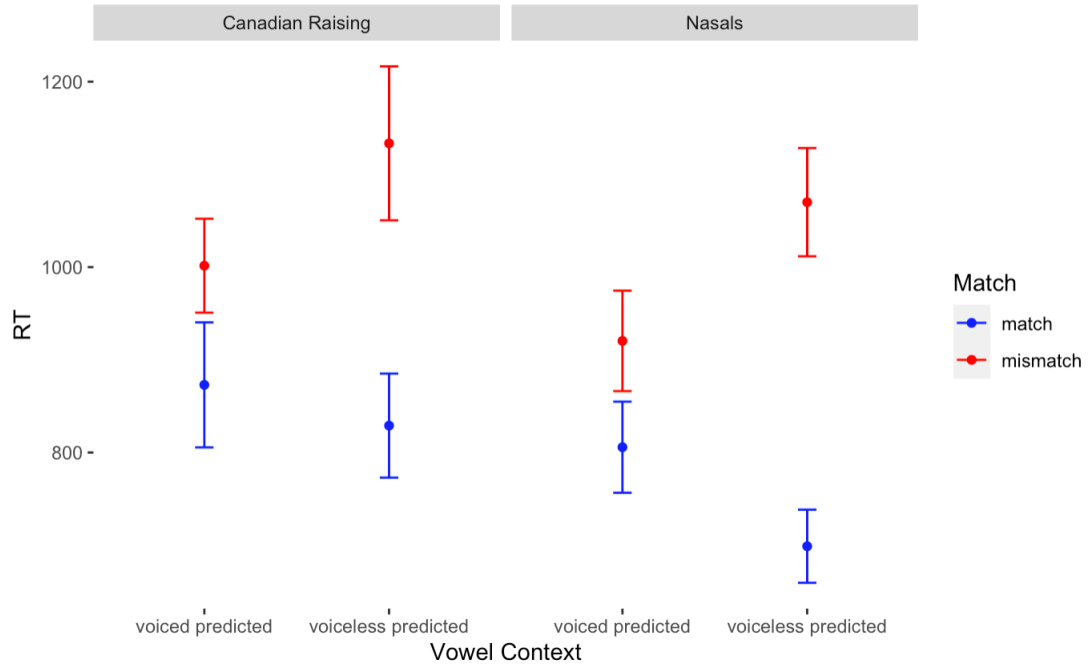

Figure 7: Exp 2 RT (ms) for each condition, with vowel marked by prediction and consonant marked by match

The response time data showed a similar pattern for both conditions, and one that suggests an allophonic representation. In the "voiced predicted" cases, where the vowel should make a less strong prediction, there appeared to be a small effect of match. The voiced consonant had a faster response, and there was a slight slowdown for the voiceless consonant. For the "voiceless predicted" cases, however, there was a different pattern. The voiceless consonants had a speed-up and had the fastest response times overall, especially with the nasals. The voiced consonants had a steeper slowdown and were the overall slowest.

4.2.3 Modelling

In addition to a general model, I modelled the data with k-folds cross validation with a *k* of 8. I partitioned the data randomly into 8 parts and ran 8 mixed effects linear regression models using the *lmer()* function, each with 1 of the 8 parts held out. Using this type of validating is a way to confirm the effects are consistent across the dataset and not coming from some small section of extreme data.

Response times are the primary measure of interest in this experiment and so I do not present an analysis of accuracy here. Fixed effects were prediction (vocalic context) and match. Random effects were subject and item. Match was contrast coded with match coded as -.5 and mismatch coded as .5. Prediction/vocalic context was contrast coded as voiceless

predicted = .5 and voiced predicted = -.5. All response times are log-transformed.

The model before cross-validation is presented below. This includes all of the data.

Table 4: Mixed-Effects Model Parameter Estimates, Canadian Raising

	Estimate	Standard Error	t value	p value
(Intercept)	6.634	0.053	119.9811	<0.001***
Prediction	0.027	0.031	0.876	0.383
Match	0.209	0.026	8.196	<0.001***
Prediction:Match	0.130	0.051	2.571	0.010*

Table 5: Mixed-Effects Model Significances from Cross-Validation, Canadian Raising

Model	Intercept	Prediction	Match	Prediction:Match
1	<0.001***	0.3218	<0.001***	0.0118*
2	<0.001***	0.4970	<0.001***	0.0021**
3	<0.001***	0.5132	<0.001***	0.0134*
4	<0.001***	0.6874	<0.001***	0.0054**
5	<0.001***	0.6756	<0.001***	0.0257*
6	<0.001***	0.3281	<0.001***	0.0027**
7	<0.001***	0.303	<0.001***	0.017*
8	<0.001***	0.6130	<0.001***	0.0073**

For the nasal data, the exploratory analysis did not reveal the same issues with significance. One model containing all of the data is presented below.

Table 6: Mixed-Effects Model Parameter Estimates, Nasals

	Estimate	Standard Error	t value	p value
(Intercept)	6.595	0.467	141.311	<0.001***
Prediction	0.061	0.030	2.040	0.043*
Match	0.288	0.024	12.136	< 0.001***
Prediction:Match	0.293	0.476	6.148	<0.001 ***

For both the Canadian Raising and the nasal conditions, there was a significant positive effect of match. With match coded as -0.5, this shows a reduction in response time for matches as compared to mismatches. For the nasals but not the Canadian Raising stimuli,

there a small positive effect of prediction. This predicts a small speed-up for the phonemic vowel as compared to the allophonic vowel. For the nasals, there was a significant interaction between prediction and match, which predicts a longer response time for mismatches with the allophonic vowel. This effect also reached significance in the general Canadian Raising model as well as all 8 of the cross-validated models.

4.3 Discussion

The Nasal stimuli show a higher response time for mismatches from the allophone than from the phoneme, confirming the overall hypothesis of allophonic behavior in the task. The results suggest an abstract, allophonic representation of the vowels involved in Canadian Raising. Figure 8 showed that listeners were fastest at responding to raised vowel + voiceless consonant combinations. They were faster at those matches than when they hear a raised vowel + voiced consonant. This can be explained under an allophonic account because the raised vowel is making a stronger prediction. Moreover, the participants were slower at responding to mismatches with a raised vowel + voiced consonant as opposed to mismatches with an unraised vowel + voiceless consonant. Upon hearing a raised vowel, listeners seemed to be making a strong prediction about the voicing of the upcoming consonant, and when that prediction was not met, this results in a slowdown. They did not experience slowdowns as severe with the unraised vowel, which aligns with the theory that this vowel isn't making as strong of a prediction. Importantly, the Canadian Raising stimuli showed the same overall pattern of behavior as the nasals, which further evidences an allophonic account. The magnitude of the effect did appear to be stronger for the nasal stimuli, but the direction was the same.

The accuracy results are interesting. Although the primary measure was response time, the accuracy results were not entirely surprising, Participants made more errors on mismatches than matches. For the nasals in particular, there is very low accuracy when a voiceless consonant is predicted followed by a voiced stop. It is puzzling that there is not the same asymmetry with the Canadian Raising stimuli. From Figure 6, however, it is clear that the asymmetry is most present in the 'both' and 'only voiceless' conditions. In the 'only voiceless' condition, this error means that they are responding with the voicing that makes a word, which makes sense. In the 'both' condition, they appeared to be biased towards responding with the voiceless word even when hearing a voiced stop. This is similar to the pattern with the accuracy in the gating study, where participants were strongly biased towards a voiceless response in the 'both' condition.

5 General Discussion

In experiment 1, participants had higher accuracy at earlier gates for stimuli containing allophones in both the nasal and Canadian Raising conditions. In experiment 2, participants showed slowest reaction time for mismatched cross-spliced stimuli when the vowel was an allophone and the consonant was a voiced consonant. They were also fastest for stimuli which were matches containing an allophone and a voiceless consonant.

This pattern of results suggests the allophonic representation of the vowels in Canadian Raising is the correct representation. This is based on the assumption that an allophone's restricted environment makes a stronger prediction about upcoming material than the underlying phoneme. An alternate explanation for these facts might appeal to the surface distribution and say that the prediction is not found in the phonology, but in some statistical knowledge of the lexicon as a whole and needn't be abstract.

To test this, I ran 2 bigram models of English to see if certain consonants are more predictable after these diphthongs based on the statistical distribution of the language. I ran simple bigram models, frequency weighted by the SUBTLEX frequency. The data modelled is the SUBTLEXus (Brysbaert and New 2009) corpus of subtitles and the transcriptions come from the CMU pronouncing dictionary. The CMU dictionary is phonemically transcribed, so the first model simply looks at the distribution of phonemes that follow /aɪ/. In the experiment, participants had a forced choice between two consonants at the same place of articulation. As a direct comparison, in the phonemic bigram, I compare the probabilities of pairs of consonants. In the table below, the consonant with higher predictability at a certain place of articulation is bolded. In addition, I also manually raised the vowels in the CMU transcriptions as well as flapped the /t, d/s in appropriate environments and ran a second bigram model over phones (phonetic). In this bigram model, the comparison between consonant pairs is not drawn from the same distribution as the voiceless sounds only occur after [ʌɪ] and the voiced only after [aɪ]. This bigram is more accurate to the phonetics, but a less reliable direct comparison of probabilities as they are drawn from different distributions.

Table 7: Bigram Model Results

Phonemic		Phonetic	
p(t|aɪ)	**0.199**	p(t|ʌɪ)	**0.495**
p(d|aɪ)	0.064	p(d|aɪ)	0.092
p(f|aɪ)	0.028	p(f|ʌɪ)	0.077
p(v|aɪ)	**0.188**	p(v|aɪ)	**0.300**
p(p|aɪ)	**0.005**	p(p|ʌɪ)	**0.013**
p(b|aɪ)	0.003	p(b|aɪ)	0.005
p(s|aɪ)	0.037	p(s|ʌɪ)	**0.099**
p(z|aɪ)	**0.049**	p(z|aɪ)	0.078

Both bigrams show that voiceless [t, p] are more predictable than [d, b] and voiced [v] is more predictable than [f]. Each model picks out a different phone for [s, z] as more predictable, but in each the probabilities for each are quite similar. Thus, under a purely statistical predictability model, the stronger prediction would be from [ʌɪ] for [t/d, p/b] and from [aɪ] for [f/v]. For [t/d, p/b], I would expect mismatches following [ʌɪ] to be slower than following [aɪ]. The opposite is true for [f/v], where mismatches following [aɪ] should be slower than those following [ʌɪ] as the voiced stop is more predictable. The response times from Experiment 2 broken down into place of articulation are shown in Figure 8.

As is seen in Figure 8, all the consonants followed the same general pattern within mismatches. Mismatches from [ʌɪ], labelled as "voiceless predicted" are slower than voiced predicted. Within match, there was more variation which does seem to align with the bigram models. In B, D, the voiceless predicted was responded to faster, whereas in V, the voiced predicted was faster, and in Z, there doesn't appear to be much of a difference. The matches seem to align with the bigram predictions, but the asymmetries in the mistmatches cannot be explained by a simple statistical model.

The asymmetrical pattern in the mismatch data suggests a difference between the vowels [aɪ, ʌɪ] which cannot be explained by surface patterns. The results align with the analysis that /aɪ/ is a phoneme which raises to [ʌɪ] in certain phonological contexts. The hypothesis proposed in this paper is that allophones make a stronger prediction about upcoming material due to their restrictions on distribution at a grammatical level. An alternate way to consider the facts is about matching and mismatching. If [ʌɪ] is an allophone of [aɪ], hearing an [aɪ] in a raising context is matching the vowel at a phonemic level, but not a

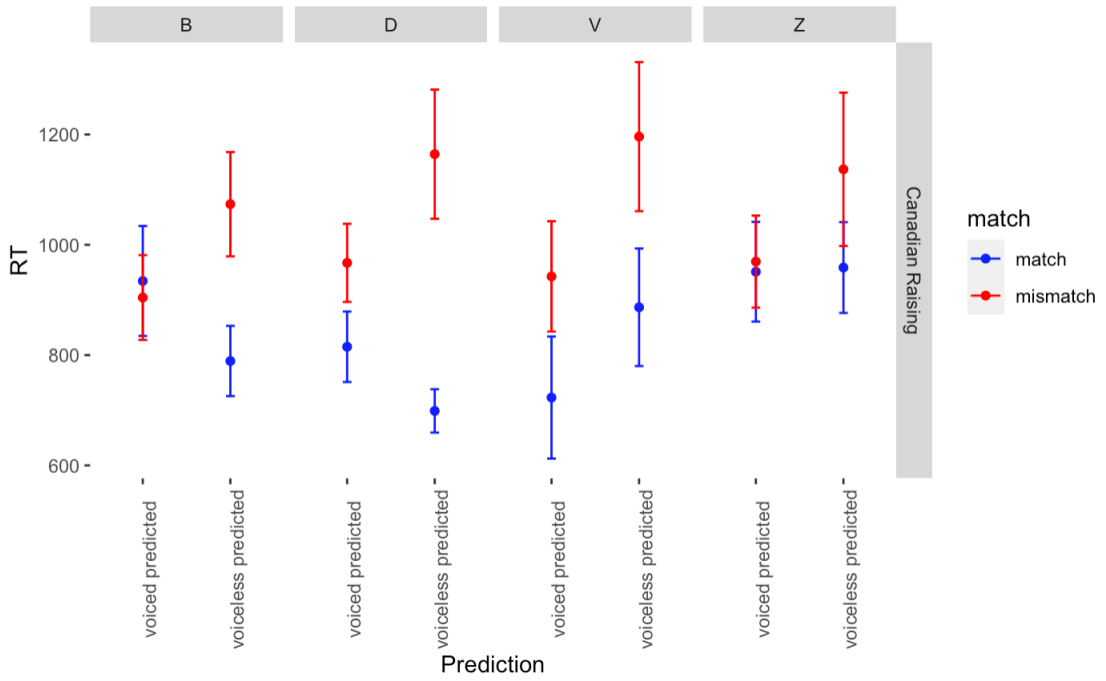

Figure 8: Exp 2 RT (ms) for Canadian Raising stimuli divided by place of articulation of final consonant. Vowel quality marked by prediction and consonant voicing marked by match.

surface level. On the other hand, hearing a [ʌɪ] in a non-raising environment is hearing a phone which is incorrect at the surface level and also does not match a the phoneme. This explanation still relies on an opaque, abstract account of these vowels, but does not rest on a prediction account necessarily. Teasing apart these two explanations is beyond the scope of this study, but what I have shown here is a pattern of results which suggests an abstract, allophonic account of the vowels involved in Canadian Raising.

6 Conclusion

This paper has presented two experiments that show evidence for the vowels in Canadian Raising being allophones of the same phoneme. This evidence is in line with the previous literature, namely Farris-Trimble and Tessier (2019). These results provide experimental support for phonological analyses of this process which assume no underlying contrast.

References

Bermúdez-Otero, R. (2004, November). *Raising and flapping in Canadian English: grammar and acquisition.* Handout of paper presented at the CASTL Colloquium, University of Tromsø.

Boersma, P., & Weenik, D. (2001). Praat: a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341-345.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavioral Research Methods*, *41*(4), 977-990. doi: 10.3758/BRM.41.4.977.

Cohn, A. C. (1993). Nasalisation in English: Phonology or phonetics. *Phonology*, *10*(1), 43-81. doi:10.1017/S0952675700001731.

Farris-Trimble, A., & Tessier, A.-M. (2019). The effect of allophonic processes on word recognition: Eye-tracking evidence from Canadian Raising. *Language*, *95*(1), 136-160. doi:10.1353/lan.2019.0023.

Gaskell, M. G., & Marslen-Wilson, W. D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(1), 144-158. doi: 10.1037//0096-1523.22.1.144.

Gaskell, M. G., & Marslen-Wilson, W. D. (1998). Mechanisms of phonological inference in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(2), 380-396. doi: 10.1037//0096-1523.24.2.380.

Gow, D. W., Jr. (2001). Assimilation and anticipation in continuous spoken word recognition. In *Journal of memory and language* (Vol. 45, p. 133-159).

Graham, L. (2019). Production and contrastiveness of Canadian Raising in Metro-Detroit English. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th international congress of phonetic scienes* (p. 127-131). Canberra, Australia.

Hall, K. C. (2005). Defining phonological rules over lexical neighbourhoods: Evidence from canadian raising. In J. Alderete, C. hye Han, & A. Kochetov (Eds.), *Proceedings of the 24th west coast conference on formal linguistics* (p. 191-199). Cascadilla Proeceedings Project.

Harris, Z. S. (1951). *Methods in structural linguistics*. University of Chicago Press.

Hualde, J. I., Luchkina, T., & Eager, C. D. (2017). Canadian Raising in Chicagoland: The production and perception of a marginal contrast. *Journal of Phonetics*, *65*, 15-44. doi: 10.1016/j.wocn.2017.06.001.

Idsardi, W. (2006). Canadian raising, opacity, and rephonemicization. *The Cana-*

dian Journal of Linguistics, *51*(2/3), 119-126.

Joos, M. (1942). A Phonological Dilemma in Canadian English. *Language*, *18*(2), 141-144. doi: 10.2307/408979.

Krämer, M. (2019). Is vowel nasalisation phonological in English? A systematic review. *English Language and Linguistics*, *23*(2), 405-437.

McQueen, J. M., Norris, D., & Cutler, A. (1999). Lexical influence in phonetic decision making: Evidence from subcategorical mismatch. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(5), 1363-1389. doi: 10.1037/0096-1523.25.5.1363.

Mielke, J., Armstrong, M., & Hume, E. (2003). Looking through opacity. *Theoretical Linguistics*, *29*(1-2), 123-139.

Nazarov, A. (2019). Formalizing the connection between opaque and exceptionful generalization. In *Toronto Working Papaers in Linguistics* (Vol. 41).

Pater, J. (2014). Canadian raising with language-specific weighted constraints. *Language*, *90*, 230-240.

Prince, A., & Smolensky, P. (1993/2004). *Optimality theory: Constraint interaction in generative grammar*. Malden MA, and Oxford, UK: Blackwell.

Smits, R., Warner, N., McQueen, J. M., & Cutler, A. (2003). Unfolding of phonetic information over time: A database of dutch diphone perception. *Journal of the Acoustical Society of America*, *113*(1), 563-574. doi: 10.1121/1.1525287.

Zehr, J., & Schwarz, F. (2018). *PennController for Internet Based Experiments (IBEX)*.