Laboratory Phonology
Journal of the Association for Laboratory Phonology

Wade, L. 2017 The role of duration in the perception of vowel merger. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 8(1): 30, pp. 1–34, DOI: https://doi.org/10.5334/labphon.54

JOURNAL ARTICLE

The role of duration in the perception of vowel merger

Lacey Wade
Department of Linguistics, University of Pennsylvania, Philadelphia, PA, US
laceya@sas.upenn.edu

Speakers with vowel categories that are considered merged by traditional measures (e.g., F1 and F2 measurements at a single time point) may contrast vowel classes in dimensions beyond vowel quality, such as duration. Durational differences among vowel classes have been observed to persist even in cases of spectral overlap (e.g., Fridland et al., 2014; Labov & Baranowski, 2006), suggesting that duration may serve as a contrastive cue among spectrally-merged or near-merged vowel classes. This paper examines the role of duration in perception in two communities: Youngstown, OH, which exhibits multiple patterns of merger and distinction among POOL-, PULL-, and POLE-class words, and Burlington, VT, whose residents are largely unmerged. This paper presents the results of a forced-choice identification task consisting of lexical stimuli with synthetically manipulated vowel-liquid durations, analyzed in light of participants' production data. Results indicate that duration influences vowel categorization and is utilized more extensively when spectral cues are diminished or unavailable.

Keywords: vowel merger; sociophonetics; perception; production; duration

1 Introduction

Phonemic merger exists when an individual lacks a phonemic contrast.[1] Despite the prevalence of research on this topic, it is still somewhat unclear which types of empirically observable phenomena indicate presence or absence of an underlying contrast. Because there are various—and often contradictory—measures of both production and perception that might be utilized in determining whether an individual distinguishes between two phonemes, the dividing line between an unmerged and merged speaker is not easy to draw.

One problem with having various measures to choose from is empirical: Measures of production and perception—or even different measures of either production *or* perception—have been observed not to align (e.g., Herold, 1990; Labov et al., 1991, 1972; Nycz, 2013). For instance, individuals may perceive vowel contrasts that they do not produce; such findings have often been attributed to merged speakers' contact with unmerged speakers (Hay et al., 2013; Warren & Hay, 2006; Hay et al., 2006; Thomas & Hay, 2005) or the influence of orthography (Faber & Di Paolo, 1995; Herold, 1990). Alternatively, speakers may consistently produce a distinction between vowel classes that they are unable to utilize in perception, referred to as 'near-merger' (Labov et al., 1972). Somewhat less discussed in the literature is when different measures of the same aspect

[1] A lack of phonemic contrast might occur only in certain phonological environments, as is the case with conditioned mergers like the PIN-PEN merger, or in all environments, as is the case with unconditioned mergers like the low back merger.

of speech (e.g., different measures of production) fail to align. Labov et al. (1972) recount the case of Dan Jones, who produced distinctions between POOL- and PULL-class words in casual speech, primarily in the F2 dimension, though he produced no distinction when reading a list of minimal pairs. Another speaker, Bill Peters, produced a consistent low-back distinction in spontaneous speech, but this distinction was also largely diminished when reading a list of minimal pairs (Labov et al., 1972). Dan Jones' and Bill Peters' patterns have been explained as stemming from different levels of attention to linguistic form elicited by each task.

The second problem with diagnosing presence of merger is a matter of conceptual confusion. Speakers may be considered merged in production by conventional diagnostics such as measurements of F1 and F2 at a single time point within the vowel, even if vowel classes are kept distinct in production by some other dimension of the speech signal. For instance, Milroy and Harris (1980) suggest that the merger of MEAT and MATE in Belfast English is actually a near-merger because these vowel classes are produced with consistently different inglides. Even when spectral dynamics are taken into consideration, phonetic merger is often equated with phonemic merger. However, as Maguire et al. (2013), for instance, note in a squib focusing on defining merger, phonetic merger is not the same as phonemic merger. They discuss a number of non-categorical phenomena, including "inter-speaker variability, intra-speaker variability, partial merger, near merger, and merger in production but not perception," that researchers often mistakenly equate with *phonemic merger* (p. 233). It is important to note that merger in one or more acoustic dimensions does not necessarily indicate a merged underlying vowel class, as another acoustic cue may aid in discriminating between the vowel classes. Looking more closely at instances of spectral merger, which many have considered to be indicative of a single merged phonemic category, may yield evidence to the contrary; spectral merger may be phonemic, but need not be, since vowel classes have the potential to be distinguished by other acoustic cues beyond vowel quality. Spectral merger, then, is an ideal condition for beginning to probe the question of which empirically observable phenomena truly indicate lack of an underlying phonemic contrast.

Duration is one dimension in which spectrally-merged vowel classes might be contrasted. It has been well documented that English vowels have inherently different durations that correspond to both height and peripherality in the vowel space (Peterson & Lehiste, 1960; Watson & Harrington, 1990). More recently, it has also been shown that, in cases of spectral overlap, length distinctions may be maintained or even enhanced to distinguish between two otherwise merged vowel classes (e.g., Labov et al., 2006; Labov & Baranowski, 2006; Fridland et al., 2014). These findings naturally bring up questions regarding the phonemic status of length distinctions in English, as well as the extent to which production maps onto perception in cases of spectral merger. Whether speakers who make durational contrasts in production can perceive durational distinctions and utilize durational cues productively in everyday communication is an empirical question for which there is not yet a clear answer.

The present study seeks to shed light on these questions by examining the degree to which listeners in two different speech communities are influenced by duration in a vowel classification task for which the vowel-liquid sequence of each target lexical item has been synthetically manipulated. The community of interest is Youngstown, Ohio, a Rust Belt city in northeastern Ohio, which has been previously shown to exhibit multiple patterns of merger and distinction among POOL-, POLE-, and PULL-class words (Arnold, 2015). The POOL-PULL and POLE-PULL mergers have been shown to be different from one another in that only the POOL-PULL merger shows a strong production-perception relationship, and this merger is relatively stable in apparent time. The POLE-PULL merger,

on the other hand, is relatively nascent in the community and is progressing in apparent time. Youngstown listeners' perception results are compared to those of listeners from Burlington, Vermont, a community that largely maintains distinctions among pre-/l/ back vowels. Results of this perception task are analyzed in light of production data in order to investigate whether durational distinctions are more perceptually salient in the absence of vowel quality distinctions. The following research questions guide this inquiry:

1. Do spectrally-merged speakers use duration to differentiate vowel classes in production?
2. Do listeners rely on durational cues in perception, and if so, do production patterns influence use of these cues?
3. How does exposure to community patterns influence the degree to which duration is utilized as a perceptual cue?

Experiment 1 investigates how speakers with multiple patterns of merger and distinction make use of durational cues in production. Results suggest that both spectrally-merged and unmerged speakers may make use of durational cues in production, though there is considerable interspeaker variation. In light of the results from Experiment 1, Experiment 2 seeks to determine how these durational cues might be utilized in perception. The results of Experiment 2 indicate that both spectrally-merged and unmerged speakers utilize duration to some extent when identifying vowel classes, but duration is utilized most when more salient cues such as vowel quality are diminished or simply unavailable, either because 1) speakers do not produce reliable vowel-quality distinctions in their own speech, or 2) speakers cannot rely on vowel quality distinctions in perception because of inconsistent maintenance of distinctions in the community. I ultimately propose that alternative cues such as duration may play an important role in phenomena such as near-merger and thus offer valuable insights into the nature of phonological contrasts.

1.1 Duration as a cue in production and perception

Cross-linguistically, vowels with longer durations tend to be more peripheral or open (Peterson & Lehiste, 1960). This pattern can be attributed to physiological constraints; the more peripheral or open the target, the wider the mouth must open and the longer it will take to reach the target gesture. Still, durational patterns vary by language. Most obviously, many languages have phonemic length distinctions (e.g., Japanese, Arabic, Thai, German) while other languages (e.g., English) do not. Though English is not typically considered to have phonemic length, duration in English may vary due to stress, speaking rate, vowel-inherent properties, and phonological environment. It has even been suggested that vowel duration is a more important cue for signaling following consonant voicing than the consonant-internal cues (e.g., Chen, 1970; Kohler, 1979; Raphael, 1972), which might indicate an allophonic length distinction in English.

Several recent studies have confirmed that duration varies not only cross-linguistically but also cross-dialectally. Southern speakers produce longer durations for many vowel classes (Clopper et al., 2005; Fridland et al., 2014; Jacewicz et al., 2007), and duration may be employed contrastively in one dialect but not in another (Labov & Baranowski, 2006). Such findings suggest that durational distributions in English do not always stem from physiological constraints. Tauberer and Evanini (2009) for instance, found that changes in vowel openness by dialect region in North American English did not correspond to changes in duration, suggesting a stored target duration that exists independent of location in the vowel space. Similarly, Langstrof (2009) found that in the New Zealand English vowel shift, changes in vowel height did not result in corresponding shifts in duration. Accumulating evidence suggests that duration has a larger role in English than

as a consequence of articulatory constraints alone. It comes as no surprise, then, that listeners attend to duration in perception as well.

Multiple cues may be relevant for distinguishing between vowel classes, though duration has been noted as particularly salient, especially for listeners learning non-native contrasts. Listeners have been shown to rely primarily on durational cues in distinguishing between tense and lax vowel classes, for instance, regardless of whether durational cues exist in their L1. This is the case even for contrasts in which vowel quality is the primary cue (e.g., the tense-lax distinction in American English) (Bohn, 1995; Cebrian, 2006; Escudero, 2000). For instance, Cebrian (2006) found that native Catalan speakers utilized duration as the primary cue for distinguishing English vowel contrasts to a greater extent than native English speakers, who rely on duration only as a secondary cue. Though it is not clear what it is about durational cues that make them so easily accessible, Bohn (1995) suggests that listeners may rely so heavily on duration because of a "general speech perception strategy that takes over whenever information conveyed by spectral differences is insufficient" (p. 300).

Duration has also been empirically shown to influence native listener perception in several studies on English vowels, though the effects have been subtle and are often framed as being secondary to spectral cues. Sawusch and Palmer (1997), for instance, showed that duration had minimal effects on recognition of /æ/ and /ɛ/ alone, but had a much greater effect when combined with manipulations of formant movement across the duration of the vowel. Hillenbrand et al. (2000) also found that alterations in vowel duration had a relatively small effect on participants' abilities to identify vowels, as almost all were identified correctly for both the original and synthetically manipulated tokens. Spectral information, therefore, was found to be much more important. Ainsworth (1972), however, suggested that duration played a somewhat larger role in vowel identification, at least for vowels located more centrally in the vowel space which therefore have a higher chance of being confused for other vowels. Similarly, Bennet (1968) found an inverse relationship between the use of durational cues in perception and the distance in the vowel space between a given pair of vowels. Labov and Baranowski (2006) found that Inland North speakers produced a 54-millisecond difference in duration, on average, between /ɛ/ and /ɑ/, which show complete spectral overlap for these speakers. Undergraduate participants were able to distinguish between these spectrally-overlapping productions based on durational differences alone. The consensus seems to be, then, that duration is a perceptual cue that may be utilized in combination with other more salient cues or when primary cues are weak or unavailable.

1.2 The role of alternative contrastive cues in near-merger

Contrasts made in dimensions that are not typically considered contrastive in a language (e.g., duration in English) are particularly interesting in cases of near-merger—the production-perception mismatch where a speaker consistently makes subtle distinctions between vowel classes, which he or she cannot perceive. In terms of perception, there does not seem to be much difference between near-merger and complete phonemic merger, as listeners cannot reliably discriminate between vowel classes in either case, which likely complicates near-merger diagnostics.[2] As the present study argues that alternative contrasts may play a role in phenomena that are often diagnosed as near-merger, it is important to

[2] The literature on merger has not attempted to draw a distinction between the discriminatory abilities of near-merged and traditionally merged speakers, though it might be expected that near-merged listeners would perform at least somewhat better than a traditionally merged speaker on a discrimination task. After all, in order to make a distinction in production, no matter how small, some knowledge of vowel class categories is likely necessary.

understand how presence of near-merger is traditionally determined, particularly since the diagnostics often utilized in determining presence of near-merger do not target such alternative contrasts.

For instance, when diagnosing near-merger, sociolinguists have often relied on minimal pairs tasks, which are based on participants' self-judgments regarding whether they believe a minimal pair differing only by the vowel contrast of interest should be pronounced the same or different (e.g., Di Paolo & Faber, 1990; Labov et al., 1991, 1972). The advantage of such tasks is that they get at speakers' conscious knowledge of vowel productions. However, judgment tasks of this type fail to determine whether participants can categorize vowel classes correctly, perhaps using cues that they cannot consciously identify. In fact, it has been found that some individuals *can* accurately categorize vowel classes that they judge as the same in their own production (e.g., Thomas & Hay, 2005). Perceptual abilities are generally determined with commutation tests, which typically involve auditory presentation of isolated words, often from participants' own speech, which they are then asked to identify, usually in a multiple forced-choice format. Studies that use this method to diagnose near-merger often report that speakers do not 'pass' commutation tests unless they achieve 100% accuracy rates; however, many of these 'failures' are still at above-chance (e.g., 80%) levels (e.g., Bowie, 2000; Labov et al., 1991, 1972). With 100% accuracy as the cut-off point for a perceptual distinction, it is unclear whether participants who 'fail' commutation tests are still picking up on some sort of perceptual cues—perhaps cues beyond vowel quality. Further, since near-merger is most often present in communities undergoing change and therefore exhibiting patterns of both merger and distinction (and likely contradictory primary and secondary cues), it may be the case that participants do not perform at 100% accuracy rates because of either inconsistent cue weighting strategies, or consistent reliance on inconsistent cues. One aim of the present study seeks to test this explanation by determining whether speakers utilize such alternative contrasts in word identification tasks even when they conflict with spectral cues.

Though the traditional description of near-merger offered by Labov et al. (1991) involves a subtle distinction in vowel quality—most often just in the F2 dimension—that speakers cannot consciously perceive and often do not produce in careful speech, the idea that the 'subtle distinctions' in production that define near merger might go beyond vowel quality is not new (e.g., Di Paolo & Faber, 1990; Faber & Di Paolo, 1995). Such alternative contrasts have sometimes been framed as enhancements of existing peripheral cues or even innovations of cues that serve to preserve contrasts between vowel classes in instances where vowel quality has become no longer contrastive (e.g., Labov & Baranowski, 2006). Such claims are compatible with reports of cue-trading relations (e.g., Repp, 1983; Kohler, 1979; Jessen, 1998), which suggest that, when two or more acoustic cues signal the same contrast, one cue may be enhanced if the other is weak or ambiguous. Though generally applied to perception, cue-trading in production has also been examined (e.g., Howell, 1993). Another possible origin of alternative contrastive cues is that they are simply *maintained* when other cues (i.e., spectral cues) are lost. Similar proposals have been made regarding a coarticulatory source of sound change. For instance, Beddor (2009) suggests that nasal coarticulation on vowels preceding nasals may persist even when the environment that caused such a change in the vowel (i.e., presence of the following nasal segment) is lost. Alternative contrastive cues in near-merger may result from similar mechanisms.

One of the dimensions in which alternative contrastive cues might be either enhanced or maintained is in the shape of the trajectory of the vowel. For instance, Majors (2005) found that the low-back distinction is maintained by some speakers in Missouri who complete

the transition in F2 from midpoint to consonant faster for /ɔ/ than for /a/, and Milroy and Harris (1980) found that the MEAT and MATE classes in Belfast have consistently different inglides. Similarly, Bowie's (2000) study on pre-/l/ mergers in Waldorf, Maryland, also hinted that speakers may maintain subtle distinctions between /ul/ and /ʊl/ classes via production of a 'broken vowel' for the /ul/ class. Other studies of near-merger have proposed contrast maintenance in dimensions beyond formants altogether. For instance, Di Paolo and Faber (1990) found that, when distinctions are lost in vowel quality between /ʊ/ and /u/ before /l/ in Utah, vowel classes may still be distinguished by differences in phonation. In a later study, they note that pitch may also be a contrastive feature (Faber & Di Paolo, 1995).

Similarly, several studies have also found that durational differences may play a role in maintaining distinctions among qualitatively merged vowel classes. Wassink (2006), for instance, found that modeling vowel overlap in just F1 and F2 space resulted in a higher degree of overlap than if duration was added to the model, suggesting that durational distinctions play a significant role in distinguishing vowel classes. Similarly, Labov and colleagues cite the spectral overlap of monophthongized /aw/, a characteristic of Pittsburgh English, with /ʌ/, which is distinguished by length (Labov & Baranowski, 2006; Labov et al., 2006). They suggest that the length difference is not merely phonetic but actually serves to differentiate the vowel classes since the durational distributions do not overlap and the means are more than five standard deviations apart. Labov and Baranowski (2006) have even suggested that durational differences can be innovated to differentiate phonemes that do not differ in duration even in unmerged speech. Such is the case with /ɛ/ and /ɑ/, which show spectral overlap for many speakers in the Inland North. It is also worth noting that presence of durational distinctions, even for the same vowels, may vary depending on the community. Fridland et al. (2014) found that duration preserves the low back vowel distinction in Nevada, especially in cases of considerable spectral overlap between the two classes, while several other studies have failed to find significant durational distinctions between the same vowels in cases of merger (Benson et al., 2011; Irons, 2007; Majors, 2005). To my knowledge, whether durational contrasts are utilized in production or perception for the pre-lateral mergers of focus in this study has not yet been examined. The present study therefore offers new insights into the role of duration as a contrastive cue in English.

2 Experiment 1: Production

Experiment 1 lays the foundation for the perception task in Experiment 2 by first examining production patterns in the Youngstown, Ohio community. This first experiment seeks to first determine whether the POOL-PULL or POLE-PULL merger is complete in the Youngstown community, or whether different durational distributions may keep these vowel classes distinct, as well as to investigate the relationship between spectral and durational distributions in production.

2.1 Methods

2.1.1 Participants

Forty-one Youngstown, Ohio natives born 1933–2005[3,4] participated in the production study. These speakers were mostly acquaintances of the researcher and participated in

[3] Speakers from such a large age range were sought in order to be sure that both merged and unmerged speakers are represented, at least for the POLE-PULL merger, which is nascent in the community (Arnold, 2015).

[4] Birth year breakdown by decade is as follows: 1930s (N = 1), 1940s (N = 2), 1950s (N = 9), 1960s (N = 2), 1970s (N = 8), 1980s (N = 11), 1990s (N = 4), 2000s (N = 4).

the study in their home, the researcher's home, or a quiet room in a public building. All participants were white, and gender was represented evenly with 20 participants who identified as male and 21 who identified as female.

2.1.2 Data collection

Production data was collected primarily from read speech, as words of interest containing /ʊ/, /u/, and /o/ in pre-/l/ contexts are relatively rare in conversational speech, and sociolinguistic-style interviews did not elicit a sufficient number of these tokens. Though speakers tend to pay more attention to their speech when reading aloud than in conversation, using read speech allows for both sufficient numbers of tokens and control over the tokens elicited (i.e., phonological context). As mergers are typically below the level of conscious awareness, attention to speech is expected to have minimal influence on production of merger/distinction.

Participants were administered four reading tasks in an order that ranged from least formal (less attention to speech) to most formal (more attention to speech): 1) a reading passage, 2) a list of sentences, 3) a list of individual words, and 4) a list of minimal pairs. The reading tasks were designed specifically for this study to elicit stressed tokens of interest. In the list of sentences, all target items appear sentence-finally. These readings were recorded with a Marantz digital recorder. The same target tokens were elicited for most speakers, so differences in coarticulatory effects of preceding consonants should not result in consistent differences across speakers. The majority of target vowels elicited (89%) were in post-labial contexts, despite the coarticulatory effect of lowered F2 after labials, so that sufficient numbers of minimal pairs could be elicited (e.g., *bull/bowl, full/foal/fool, pull/pole/pool*). Approximately 40 tokens were elicited per speaker. A complete list is provided in the Appendix. Tokens that participants skipped, stumbled over, or clearly mispronounced (e.g., *foul* for *foal*) were omitted from the data set. Because the minimal pairs task elicited the highest number of tokens, analyses were primarily based on formal speech. However, tokens from the minimal pairs task were compared with those from the reading list for each speaker to ensure context did not significantly influence vowel quality beyond subtle vowel reduction in faster speech. Additionally, pre-lateral /o/ tokens are somewhat underrepresented because this vowel was not initially of interest and because of the common mispronunciation of the word *foal* due to unfamiliarity.

2.1.3 Analysis

Praat (Boersma & Weenick, 2016) was the primary acoustic analysis software used for all measurements. LPC settings, including maximum formant height (Hz), and number of formants were adjusted for each speaker, and occasionally for each token, as needed. No attempt was made to demarcate the boundary between vowel and coda /l/ due to the vocalic quality of coda /l/.[5] Therefore, all vowel measurements presented are actually of the entire vowel-liquid sequence.

Each token was measured with two different scripts in Praat. The first took measurements of the first three formants at the onset (25% into the vowel-liquid sequence) and the midpoint (50% into the vowel-liquid sequence), though only measurements for the 25% mark will be reported, as this is where the largest vowel quality differences between vowel classes occurred. The second script was used for vowel-trajectory measurements. This script measured the first three formants at 11 time points throughout the vowel-liquid sequence.

[5] Speakers in this region often produce vocalized /l/, which contributed to the unfeasibility of measuring vowel duration separately from the following liquid. Measurements of interest were taken at the 25% mark, so the influence of vocalized /l/ on vowel quality should be minimal.

For optimal accuracy, both scripts required hand selection of the desired segments so that settings could be adjusted before running the script. Boundaries were considered to be the onset and offset of energy at F2 and higher.

When speakers are binned into binary 'merged' and 'distinct' groups, these determinations were made using Analysis of Variance (ANOVA) models in R (2015). Formant value (F1 or F2) was given as the dependent variable and Vowel Class as the independent variable, with Duration and Passage Type as fixed effects. In this community, /ul/ generally does not merge with /ol/ except in cases of triple merger, so /ʊl/ is used as the reference level in all analyses. Separate tests were run for each formant (F1 and F2) and for each time point (25% and 50% into the vowel-liquid sequence). Since a distinction at any point throughout the trajectory of the vowel means a distinction is made between vowel classes, if a significant difference ($p < .05$) was found in either F1 or F2 between /ʊl/ and /ul/ or /ʊl and /ol/ at either the onset or midpoint marks, a speaker was classified as 'distinct.' Only if no significant difference was found between vowel classes in both F1 or F2 at both time points were speakers considered to be spectrally 'merged.' In subsequent analyses, only the onset tokens taken at 25% into the vowel-liquid sequence will be used, as this time point is where the three vowel classes diverged the most. Binary classifications of 'merged' and 'distinct' will be used only sparingly and should be interpreted with caution. These classifications are used to divide the data into those with high spectral overlap and those with low spectral overlap, but I do not use these classifications to make any claims about individual speakers' underlying representations, nor about where to draw a cut-off line between merged and unmerged speakers. When possible, Euclidean distance measures between vowel classes will be used as the measure of distance between vowel classes. Euclidean distance is measured as the distance in the vowel space (F1 and F2) between a speaker's mean values for each vowel class. For instance, the Euclidean distance between a single speaker's PULL and POOL classes would be calculated as $\sqrt{(F1_{POOL} - F1_{PULL})^2 + (F2_{POOL} - F2_{PULL})^2}$, where F1 and F2 values are the mean values for each vowel class for that speaker.

All duration measures used in the analyses were taken from read connected speech and normalized by dividing the duration of each token by the speaker's average speech rate, determined using a Praat script designed by De Jong and Wempe (2009), which calculates speech rate by automatically detecting syllable nuclei based on presence of voiced intensity peaks followed by dips in intensity. For analyses involving inter-speaker comparisons, token measurements were normalized using the Lobanov method (Lobanov, 1971) in R.

2.2 Results

Speakers in Youngstown, Ohio exhibit four patterns of merger/distinction among back vowels before /l/, illustrated in **Figure 1**: Merger of POLE and PULL, merger of POOL and PULL, a triple merger among POOL, PULL, and POLE, and distinction among all three. Notably, POOL only merges with POLE in cases of triple merger. For this reason, analyses will focus on the POOL-PULL and POLE-PULL pairs separately.

Figure 2 plots Euclidean distance against duration to explore whether spectrally-merged speakers (lower Euclidean distance) and spectrally distinct speakers (higher Euclidean distance) utilize duration in production differently. The top two facets plot the mean duration difference[6] between each vowel pair for each speaker on the y-axis. These top two facets show that there is a positive correlation between Euclidean distance

[6] Duration difference is measured as each participant's mean tense vowel duration (POOL or POLE) minus their mean duration for PULL. Positive duration differences, therefore, are in the expected direction.

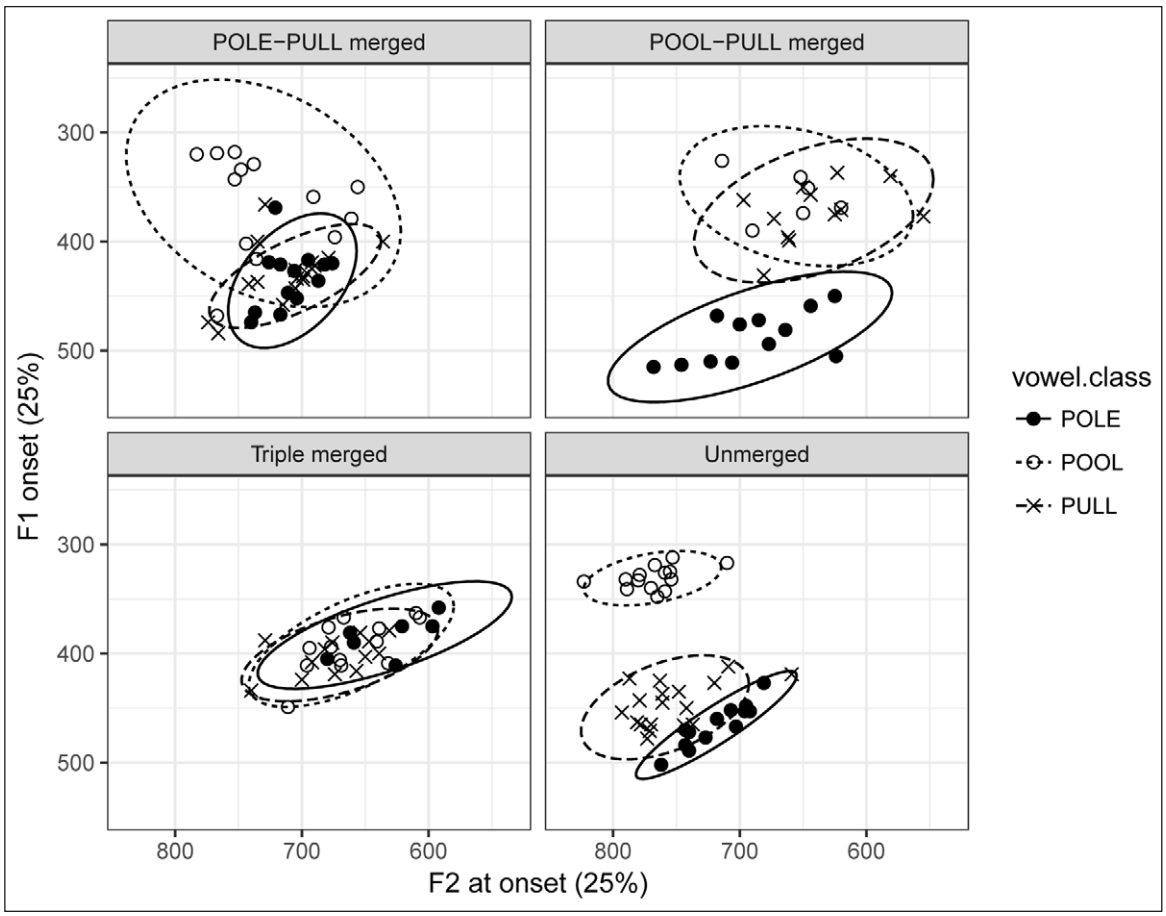

Figure 1: Possible patterns of merger/distinction in Youngstown Ohio. Each facet displays individual tokens for a single speaker who exhibits that pattern.

and duration difference for POOL and PULL (Pearson's R = .331, p = .034, but not for POLE and PULL (Pearson's R = .021, p = .904);[7] speakers who produce POOL and PULL further apart in the vowel space also produce greater durational distinctions between the two classes. Only at (log) Euclidean distances below 4 do participants produce durational distributions in the reverse of what is expected, that is, PULL longer than POOL.

The bottom two facets further elucidate the patterns shown in the top two facets by breaking down the duration *difference* seen in the top two facets into mean duration measures *for each vowel class.* The bottom right facet of **Figure 2** shows that, regardless of how far apart in the vowel space speakers produce POLE and PULL, the tendency is to maintain durational distinctions between these vowel classes. This means that many speakers who are spectrally merged may actually use durational contrasts to distinguish between these vowel classes in production, hinting that the spectral merger may not be phonemic. Though the graph shows overall longer durations for both POLE and PULL for speakers with higher Euclidean distance measures, this trend is not significant ($p > .05$) for either pair.

For the POOL-PULL pair, speakers with lower Euclidean distance measures on average do not appear to utilize duration in production (bottom left facet). Duration of PULL remains

[7] Note, though, there is much less variation in Euclidean distance between vowel classes for the POLE-PULL pair. A participant pool with a wider range of Euclidean distances for this vowel pair might show different results.

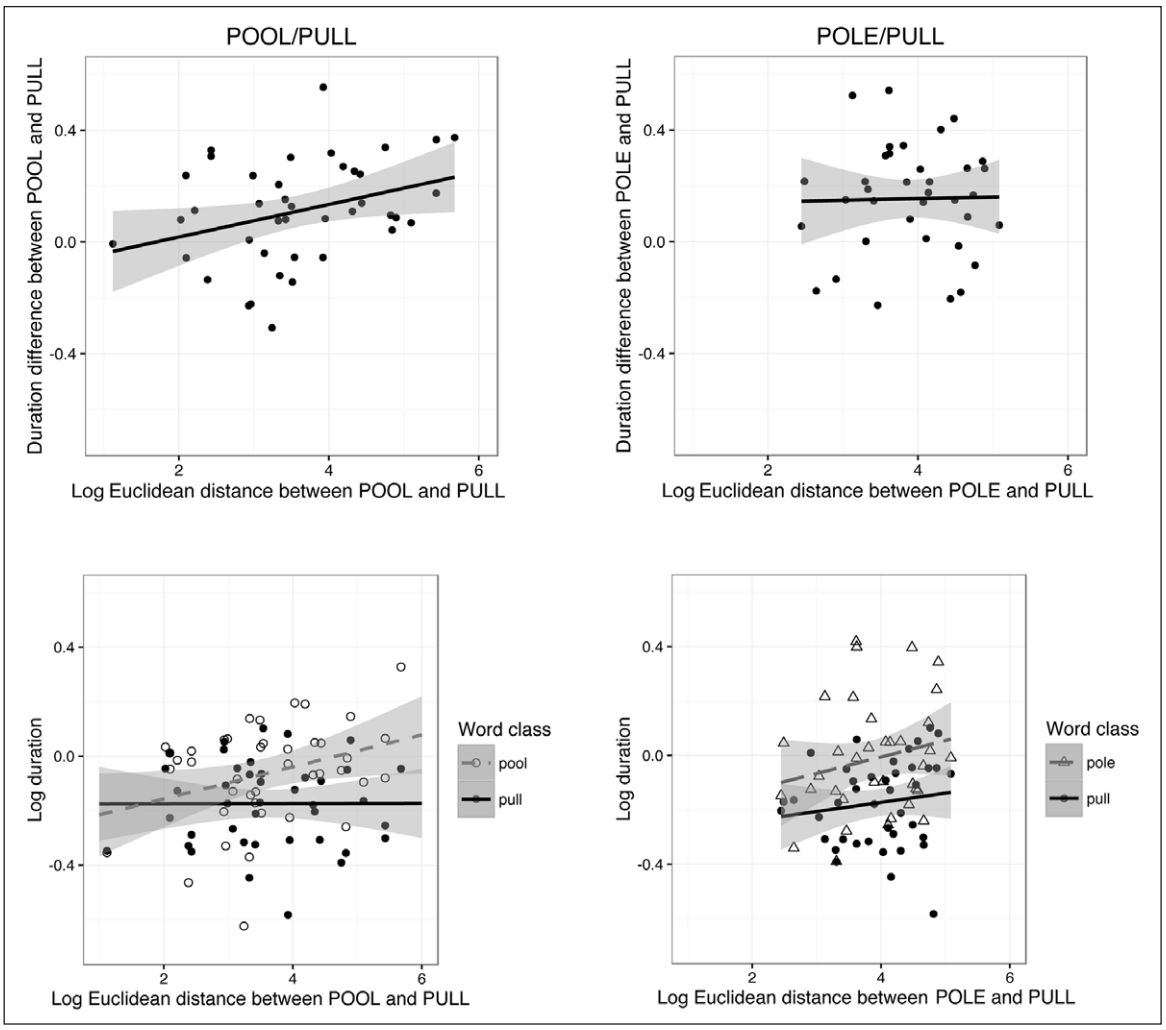

Figure 2: Correlation of duration with distance between vowel classes in the vowel space (measured in Euclidean distance between each speaker's average F1 and F2 for each vowel class). The top two graphs show the average *difference in duration* for each speaker between POOL and PULL (top left) and POLE and PULL (top right). The bottom two graphs show speaker average durations *for each vowel class*.

constant regardless of individual Euclidean distance measures, while duration of POOL positively correlates with Euclidean distance (Pearson's R = .331, p = .034). It is true that some individuals with lower Euclidean distance measures make use of durational differences in production. Individual differences in maintaining durational distinctions will be examined briefly in correlation with perception in Experiment 2.

It is worth noting that, even though many of these spectrally 'merged' speakers actually maintain durational contrasts between the vowel classes, this does not stem from vowel quality differences across the trajectory of the vowel. This is true for both vowel pairs. As **Figure 3** shows, distinction is primarily in F1 for unmerged speakers for both vowel pairs, and merged speakers appear to show no significant distinctions throughout the entire trajectory of the vowel-liquid sequence. **Figure 3** suggests, first of all, that the single time-point measures used throughout this paper (e.g., in calculating Euclidean distance between vowel classes) are sufficient, in that the 'merged' group appears to be truly merged across the entire trajectory of the vowel-liquid sequence. Secondly, spectrally 'merged' speakers who produce a durational distinction appear to be *only* producing

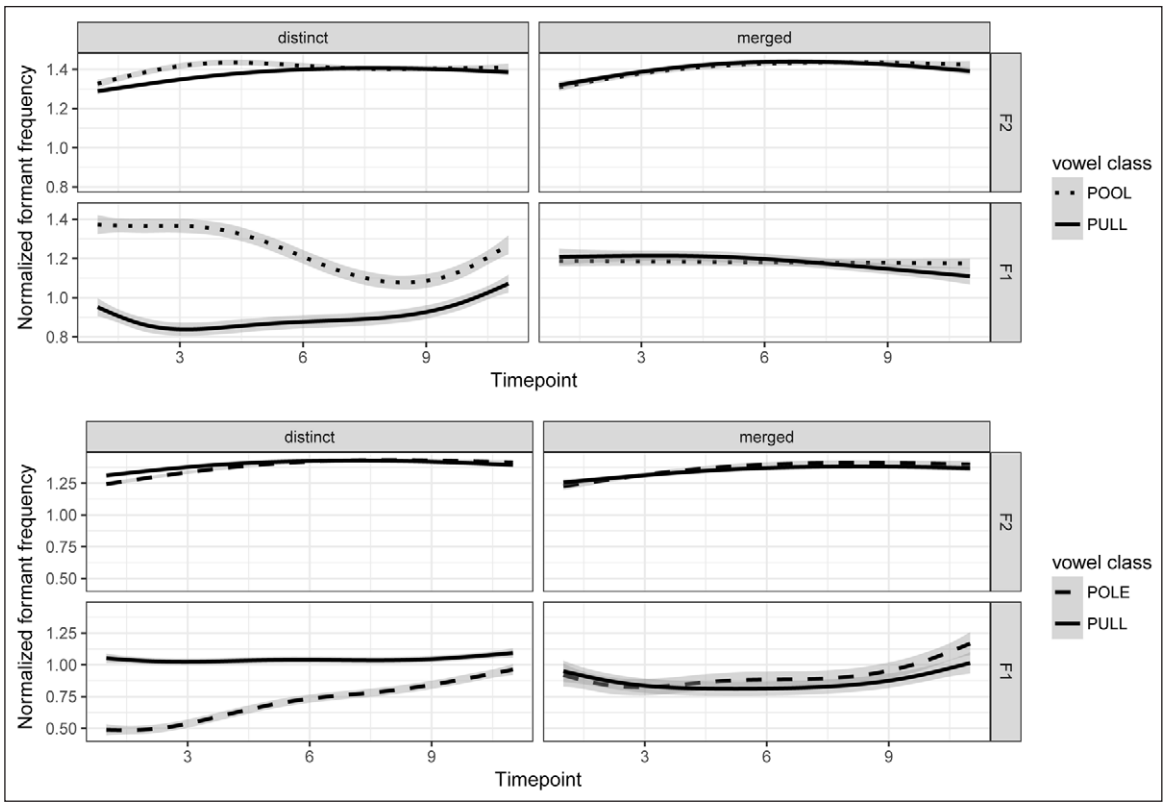

Figure 3: SSANOVA trajectories taken at 11 time points within the vowel + liquid sequence, for speakers split into 'merged' and 'distinct' categories by ANOVAs conducted for F1 and F2 at 25% into the vowel-liquid sequence.

a durational distinction. That is, durational contrasts are not somehow linked to less apparent vowel quality contrasts.

Based on production results alone, it is possible to conclude that the spectral merger of POOL and PULL is different from that of POLE and PULL. Most obviously, spectral and durational distributions correlate for the POOL-PULL pair but not for the POLE-PULL pair. This might suggest that POLE-PULL spectrally-merged speakers are not phonemically merged, as many spectrally-merged speakers utilize duration contrastively in production to the same extent as spectrally unmerged speakers. For the POOL-PULL merger, on the other hand, there is evidence that spectrally-merged speakers truly lack a contrast between these two vowel classes (i.e., the lack of trajectory differences in **Figure 3** and the correlation between spectral distinction and duration distinction in **Figure 2**).

Only by also examining how these speakers *perceive* these vowel classes can we understand the nature of this difference. For one, it is unclear with production data alone how to classify spectrally-merged speakers who produce durational distinctions. Such speakers may be near-merged in the traditional sense (i.e., unable to perceive the durational differences they make in production), or these spectrally-merged speakers may be just that—merged in vowel quality but completely unmerged (in both production and perception) in duration. The perception task comprising Experiment 2 will further shed light on these issues.

3 Experiment 2: Duration-manipulated perception task

Experiment 2 is a word-classification task comprised of auditory stimuli with manipulated vowel-liquid durations. By manipulating durational cues while holding vowel quality cues constant, it is possible to investigate not only whether participants discriminate between

either vowel pair in perception, but also which contrasts they utilize to do so. Results will be analyzed in light of participants' own production results from Experiment 1, and Youngstown participants' perception data will be compared to that of participants from a control community, Burlington, VT, in order to tease apart general perception tendencies from the influence of community-specific linguistic experience.

3.1 Methods

3.1.1 Participants

Thirty-two of the 41 Youngstown speakers who participated in Experiment 1 also participated as listeners for Experiment 2. These listeners' birth years ranged from 1947–2005,[8] and the participants were split almost evenly with 15 identifying as male and 17 identifying as female.

Sixteen participants from a control community with categorically distinct productions of back vowels before /l/—Burlington, Vermont—also participated. Burlington was chosen based on its similarity to Youngstown in many dimensions—rhoticity, presence of the low-back merger, lack of Northern Cities Shift—and distinction from Youngstown in one important area: Complete distinction of pre-lateral back vowels (Labov et al., 2006). Burlington listeners were all originally from the Burlington area and were mostly undergraduate students at the University of Vermont, though one university professor in a department other than linguistics did participate. Gender is represented evenly with 8 participants identifying as male and 8 identifying as female.

To ensure that Burlington listeners were truly not merged, production data was collected from 11 of the 16 Burlington participants. ANOVA tests confirm that each participant produced significant distinctions ($p < .05$) for both vowel-class pairs (POOL-PULL and POLE-PULL). A vowel plot of a typical Burlington speaker is shown in **Figure 4**. Further, to ensure Burlington speakers produce durational differences between vowel classes, as would be assumed if they produce spectral distinctions, a linear model with duration as the dependent variable and Vowel Class and Passage Type as fixed predictors tested this assumption. Burlington speakers produce both POOL and POLE significantly longer

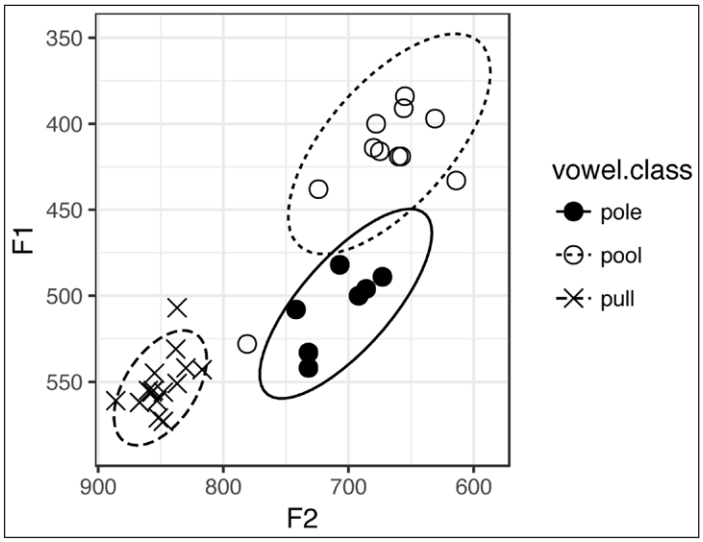

Figure 4: Vowel space of a typical Burlington speaker (male).

[8] Birth year breakdown by decade is as follows: 1940s (N = 2), 1950s (N = 7), 1960s (N = 2), 1970s (N = 7), 1980s (N = 8), 1990s (N = 3), 2000s (N = 3).

than PULL ($p < .0001$). The remaining 5 Burlington listeners participated in the study remotely through Experigen (Becker & Levine, 2014), a platform for creating perception experiments online, so their production data was not collected. A simple homophone judgment task was given to all participants to be sure they did not judge any of the vowel classes of interest as homophonous; none did.

3.1.2 Design

Twenty-six target tokens from the Youngstown speakers' production data collected in Experiment 1 were used as stimuli for Experiment 2. Target stimuli consisted only of the minimal pairs *bull-bowl*, *pull-pole-pool*, and *full-fool*.

Of the original twenty-six stimuli, nine were of the /ul/ class, ten were of the /ʊl/ class, and seven were of the /ol/ class. Each original target stimulus was manipulated to produce four new stimuli with one of four set vowel-liquid sequence durations, totaling 104 target stimuli. Forty-six filler stimuli were also included, bringing the total number of stimuli to 150. Fillers consisted of minimal pairs (e.g., *lake/Luke, pin/pen, put/pot*), also taken from the production data elicited in Experiment 1. The duration of the filler tokens was not manipulated.

The four set vowel-liquid durations were determined based on means of all connected-speech /ʊl/-/ul/-/ol/ tokens from Youngstown speakers, collected in Experiment 1. The longest duration was set to three standard deviations above the mean (383 ms), the middle longest to 1.5 standard deviations above the mean (288.5 ms), the middle shortest to the mean (194 ms), and the shortest to 1.5 standard deviations below the mean (99.5 ms). These target durations allow for a sufficient range of token durations that would mirror those of naturally occurring tokens, with the highest duration tokens being comparable to those spoken in unconnected read speech.

Vowel-liquid sequence duration was manipulated using the duration tier in Praat, which allows for lengthening or shortening of relative duration while maintaining pitch, using linear interpolation between duration points. All stimuli were adjusted using the Amplify effect in Audacity, so that each stimulus would have a peak amplitude of 1 decibel. Additionally, the Noise Reduction effect in Audacity was used to reduce background noise for all stimuli.

Stimuli from both merged and distinct speakers were used, though they were analyzed separately in order to determine the relative role of duration in processing merged vs. unmerged speech. For the distinct speakers, individual vowel plots of the most distinct speakers were examined for tokens most representative of 'distinct' tokens. Merged speaker vowel plots were also visually inspected for the most representative 'merged' tokens. In total, 10 of the original 26 tokens were taken from merged speakers and considered to be 'ambiguous,' while 16 were taken from distinct speakers and considered 'distinct.' **Table 1** shows the breakdown of which words were used as stimuli, and how they were distributed by response choice options and ambiguity.

Table 1: Breakdown of tokens used as stimuli in the perception task by ambiguity and forced-choice response options. Each of these tokens was manipulated to produce four new tokens with new vowel-liquid durations.

	POLE/PULL	**POOL/PULL**	
ambiguous	BOWL (1) PULL (1) POLE (2)	FOOL (2) FULL (1) POOL (1) PULL (2)	**10**
distinct	BOWL (1) BULL (1) PULL (1) POLE (3)	FOOL (4) FULL (3) POOL (2) PULL (1)	**16**
	10	**16**	

Stimuli were also taken from both connected (N = 9) and unconnected speech (N = 17); these sets of tokens were not analyzed as separate conditions but were chosen to balance the direction of the duration manipulation. That is, connected speech tokens were shorter and had to be mostly lengthened to conform to the four preset duration categories. Isolated tokens were longer and had to be mostly shortened to conform to the four preset duration categories.

In addition to meeting the criteria in **Table 1**, tokens were chosen based on the quality of the recording, since many were field recordings and had background noise. Aside from these restrictions, tokens were chosen at random from the speech recordings collected in Experiment 1.

The perception task was conducted using a multiple-forced choice experiment run in Praat when administered in person or the online Experigen platform when administered remotely, which was the case for 5 of the Burlington listeners. Since POLE and POOL only merge in cases of triple merger, participants were always asked to choose between either PULL and POOL or PULL and POLE.

3.1.3 Analysis

Perception task responses were coded as accurate (1) or inaccurate (0). Two logistic regression models[9] were fit to the Youngstown data set, using the lme4 package in R, one for POLE-PULL stimuli and one for POOL-PULL stimuli.[10] In each model, ambiguity is sum coded and word class is treatment coded with POLE or POOL as the reference level. Participants' mean Euclidean distance between vowel classes and mean duration difference between vowel classes in production are both centered. Duration is treated as a continuous predictor centered at the shortest stimulus duration of 99.5 ms. Two additional models were fit to both the full data set for both Burlington and Youngstown data,[11] with separate models again for POLE-PULL stimuli and POOL-PULL stimuli. Community is treatment coded with Burlington as the reference level.

3.2 Predictions

If duration is a salient perceptual cue, participants should have higher accuracy rates when duration aligns with their expectations for a given vowel class. That is, because tense vowels /o/ and /u/ are intrinsically longer than the lax vowel /ʊ/, participants should categorize /o/ and /u/ more accurately at longer durations, but /ʊ/ should be categorized more accurately at shorter durations. If participants are influenced by duration, there should be a significant interaction between duration and word class. Conversely, if vowel quality is the primary cue used to distinguish between vowel classes, participants should have high accuracy rates across the board, at least for unambiguous tokens. Slight declines in accuracy might be found for shorter stimuli, since they may be more difficult to process since listening time is reduced.

The ambiguity of the token may also play a role. If a cue trading relationship exists, it is expected that duration would be a more influential cue when vowel quality information

[9] Mixed effects models with random by-speaker and by-item intercepts were also run. Estimates and p-values were comparable to the traditional glm model, though the mixed-effects models did not converge. Results reported here are from the more stable traditional models.

[10] Two separate models were fit to the data for several reasons, both theoretical and practical. First of all, Euclidean distance and duration difference predictors are specific to one or the other vowel pair, so it makes sense to include them in only the relevant models. Secondly, this was done in an attempt to avoid empty word class cells, which would have resulted from including response options in the model. For instance, POOL-PULL stimuli would have no POLE tokens. Finally, this was done in an attempt to prevent relying on nearly uninterpretable 4- and 5-way interactions.

[11] Burlington-Youngstown comparisons must be made separately, as opposed to included in the two previously described models, because production data was not collected from roughly a third of the Burlington speakers. Any model with Burlington participants therefore cannot contain Euclidean distance or duration difference, both of which are production measures important to the analysis of Youngstown data.

is weak, such as with ambiguous stimuli. However, if durational cues are similarly influential for ambiguous and distinct stimuli, this might indicate that listeners use duration because it is an easily accessible cue, regardless of whether spectral cues are also available.

Participants' own production patterns may also influence their performance on the perception task. Experiment 1 provided evidence that spectrally-merged Youngstown speakers maintain a durational distinction for the POLE-PULL pair but not for the POOL-PULL pair. In light of these production results, spectrally-merged speakers might be expected to behave differently when listening to POLE-PULL stimuli vs. POOL-PULL stimuli, suggesting that listeners integrate their own production strategies into their perceptual behaviors. Similarly, participants who utilized duration extensively in production may also be expected to do so in perception. This would provide evidence against true near-merger and instead in favor of completely unmerged vowel classes maintained by duration rather than vowel quality. However, if listeners utilize duration across the board, regardless of their own merged status or duration usage in production, this might imply a general tendency toward relying on duration or a larger role for the speech community in influencing cue weighting strategies.

The link between listeners' linguistic experience and their own linguistic behavior is not a trivial one. Even though their phonological systems may be the same, unmerged speakers in a largely unmerged community might be expected to behave differently than unmerged speakers in a variably merged community. Exposure to a variety of vowel systems on a daily basis may force speakers to rely on only the most stable cues, or abandon acoustic cues altogether. If Burlington and Youngstown listeners behave differently, this would provide evidence for a larger role of the speech community.

3.3 Results

3.3.1 Youngstown listeners

Youngstown participants in the aggregate are influenced by duration when discriminating between not only the tense-lax pre-lateral vowel contrast (POOL-PULL), but also between the featurally distinct vowel pair (POLE-PULL). The duration effect is gradient; as **Figure 5** shows, with each decrease in stimulus duration, accuracy for POOL and POLE decreases, while accuracy for PULL steadily increases. Logistic regression models fit to the Youngstown data set (**Tables 2** and **3**) confirm that duration significantly correlates with accuracy for all word classes. In both models, there is a significant main effect of duration ($p < .001$), as well as a significant interaction between duration and word class ($p < .001$).

The duration effect in the aggregate holds up across both tokens from merged speakers (referred to here as ambiguous tokens) and those from unmerged speakers (referred to here as distinct or unambiguous tokens). As **Figure 6** shows, accuracy rates, overall, are expectedly much lower for the ambiguous tokens. In total, ambiguous tokens are recognized accurately 46.8% of the time, close to chance levels, while distinct tokens are accurately identified 75.2% of the time. Both models show a main effect of ambiguity, such that accuracy is overall lower for ambiguous tokens ($p < .001$). However, listeners do not appear to consistently utilize duration to a greater extent for ambiguous tokens, contrary to predictions. In fact, both models show a significant interaction between duration and ambiguity, but listeners are acutally influenced by duration *less* for ambiguous POOL and POLE tokens ($p < .001$). There is a significant three-way interaction in the predicted direction between duration, ambiguity, and word class for only the POLE-PULL pair, suggesting that listeners do utilize duration to a greater extent for ambiguous PULL tokens ($p = .01$). The influence of ambiguity on the extent to which listeners utilize durational cues is inconsistent at best.

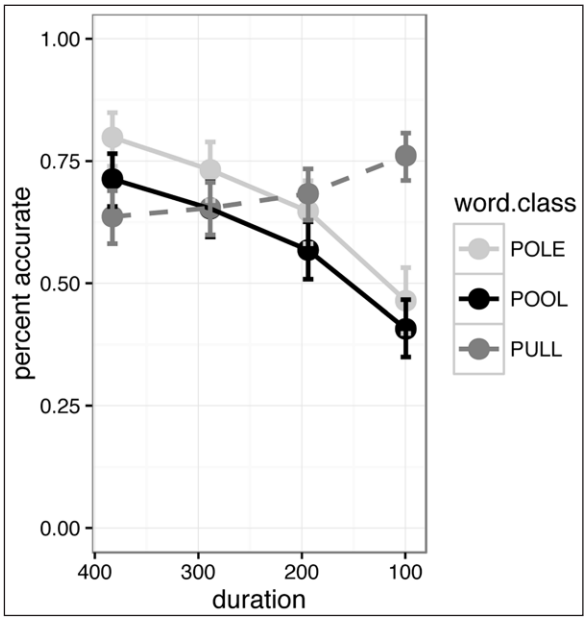

Figure 5: Aggregate Youngstown participants' accuracy for each duration category (95% confidence intervals).

Table 2: POLE-PULL logistic regression model for Youngstown listeners. Ambiguity is sum-coded while word class is treatment coded with POLE as the reference level.

	Estimate	*SE*	*z* value	Pr (>|*z*|)
(Intercept)	−0.1280	0.0924	−1.39	0.1657
Duration	0.0059	0.0006	9.71	<0.0001***
Ambiguity	−0.5056	0.0924	−5.47	<0.0001***
WordClass	0.9268	0.1792	5.17	<0.0001***
EuclideanDist	−0.0004	0.0022	−0.19	0.8479
DurationDiff	0.7832	0.4561	1.72	0.0859
Duration × Ambiguity	−0.0021	0.0006	−3.55	0.0004***
Duration × WordClass	−0.0069	0.0010	−6.59	<0.0001***
Ambiguity × WordClass	0.6431	0.1783	3.61	0.0003***
Duration × EuclideanDist	0.0000	0.0000	1.56	0.1194
WordClass × EuclideanDist	0.0072	0.0041	1.73	0.0839
Duration × DurationDiff	−0.0036	0.0029	−1.28	0.2020
WordClass × DurationDiff	−1.2696	0.8411	−1.51	0.1312
Duration × Ambiguity × WordClass	0.0027	0.0010	2.56	0.0105*
Duration × WordClass × EuclideanDist	−0.0000	0.0000	−1.82	0.0685
Duration × WordClass × DurationDiff	0.0023	0.0049	0.46	0.6432

3.3.1.1 POLE vs. PULL stimuli

This section focuses on stimuli for which participants chose between POLE-class words and PULL-class words and explains some of the unpredicted patterns displayed in the top two facets of **Figure 6**, which breaks down accuracy by word pair and ambiguity. As illustrated in these top two facets, average accuracy for ambiguous tokens are above chance levels for the longer duration tokens (evidenced by 95% confidence intervals

Table 3: POOL-PULL logistic regression model for Youngstown listeners. Ambiguity is sum-coded while word class is treatment coded with POOL as the reference level.

	Estimate	*SE*	*z* value	Pr (>|z|)
(Intercept)	−0.4799	0.0829	−5.79	<0.0001***
Duration	0.0047	0.0005	9.83	<0.0001***
Ambiguity	−0.5405	0.0832	−6.50	<0.0001***
WordClass	1.7092	0.1317	12.97	<0.0001***
EuclideanDist	0.0056	0.0012	4.64	<0.0001***
DurationDiff	0.1969	0.5324	0.37	0.7115
Duration × Ambiguity	−0.0016	0.0005	−3.38	0.0007***
Duration × WordClass	−0.0070	0.0007	−9.37	<0.0001***
Ambiguity × WordClass	−0.1868	0.1317	−1.42	0.1561
Duration × EuclideanDist	−0.0000	0.0000	−3.61	0.0003***
WordClass × EuclideanDist	−0.0097	0.0019	−5.05	<0.0001***
Duration × DurationDiff	0.0017	0.0032	0.52	0.6020
WordClass × DurationDiff	0.6315	0.8818	0.72	0.4739
Duration × Ambiguity × WordClass	−0.0002	0.0007	−0.21	0.8364
Duration × WordClass × EuclideanDist	0.0000	0.0000	4.44	<0.0001***
Duration × WordClass × DurationDiff	−0.0093	0.0051	−1.83	0.0676.

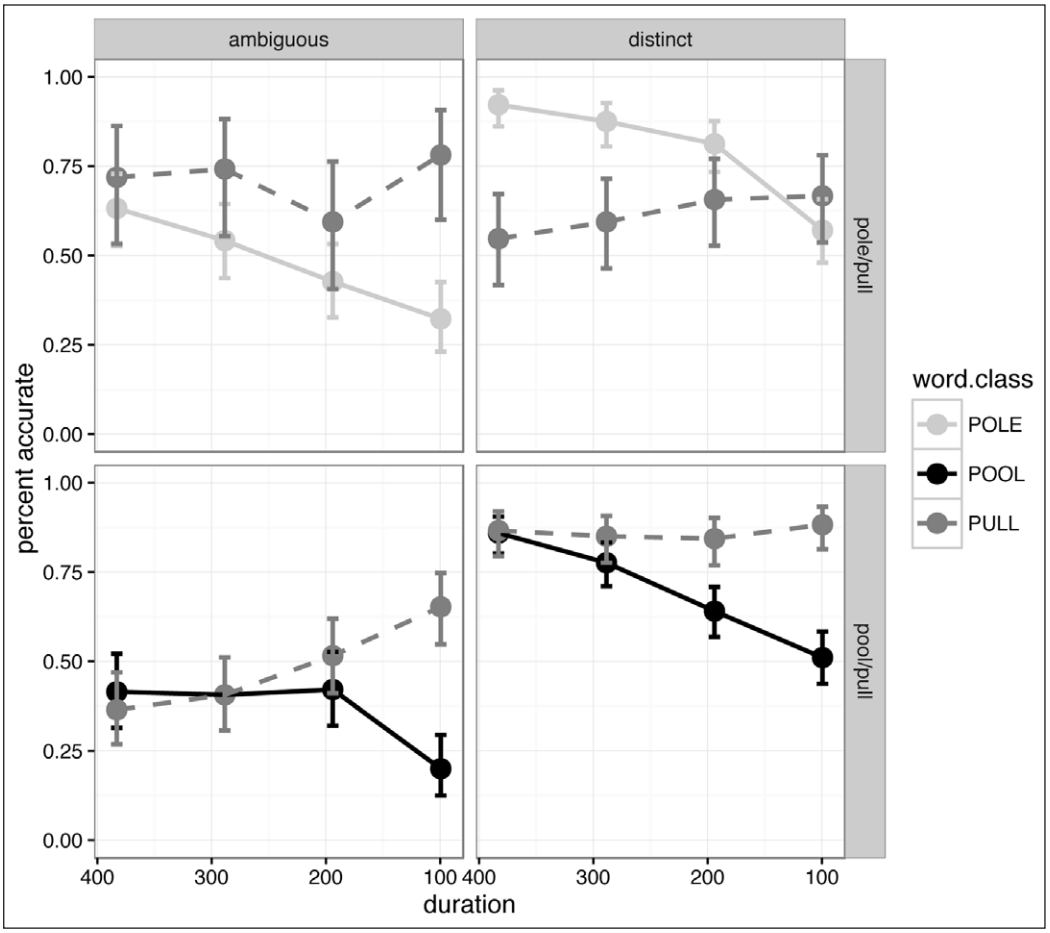

Figure 6: Youngstown participants' accuracy rates by ambiguity of stimuli and response options given (95% confidence intervals).

that do not overlap with the 50% mark), suggesting perhaps that tokens were not truly ambiguous. Though ambiguous tokens were taken from merged speakers who appear not to differentiate between PULL and POLE at any point in the trajectory of the vowel-liquid sequence, it is possible that some other dimension beyond formant values altogether may keep these two classes distinct or that the chosen ambiguous PULL-POLE tokens were somehow anomalous.

Notably, accuracy for the longest-duration distinct PULL tokens is not much different from the same ambiguous tokens, evidenced by overlapping 95% confidence intervals, yet accuracy for the longest-duration distinct POLE tokens is much higher than the same ambiguous tokens. This discrepancy in accuracy between distinct POLE and PULL can partially be explained by listener expectations: POLE tokens are expected to be longer, so accuracy is higher when this is the case. Alternatively, PULL tokens are expected to be shorter, so accuracy is lower when this is not the case. This, however, does not explain why PULL tokens that are unambiguous would be recognized at chance levels while those that are ambiguous are recognized significantly above chance levels. Recall that unambiguous tokens were produced by unmerged speakers while ambiguous tokens were produced by merged speakers. We could speculate that unambiguous PULL tokens might sound more like POOL than ambiguous tokens do because of the different realizations of PULL by merged and unmerged speakers. As **Figure 7** shows, the PULL-POLE merger is realized between unmerged PULL and POLE, meaning that the merged realization is further in the vowel space from POOL than the unmerged realization. If participants heard long distinct PULL tokens as POOL but were not given the option to choose POOL, they may have chosen an answer at random, thus explaining the near-chance accuracy levels. In fact, one participant did note that several of the sound clips for which he was asked to choose between POLE and PULL did sound more like POOL.

Shorter POLE tokens drop drastically in accuracy, the distinct tokens falling to accuracy not much different from chance, while ambiguous tokens actually fall somewhat below chance levels. This means that, for shorter duration tokens, ambiguous POLE is recognized as PULL more often than it is recognized as POLE. This is unsurprising, suggesting that participants rely more heavily on duration than vowel quality for the tokens that provide the least vowel quality information.

3.3.1.2 POOL vs. PULL stimuli

This section will turn the attention toward stimuli for which participants were asked to choose between POOL and PULL, focusing on the bottom two facets of **Figure 6**. Longest-

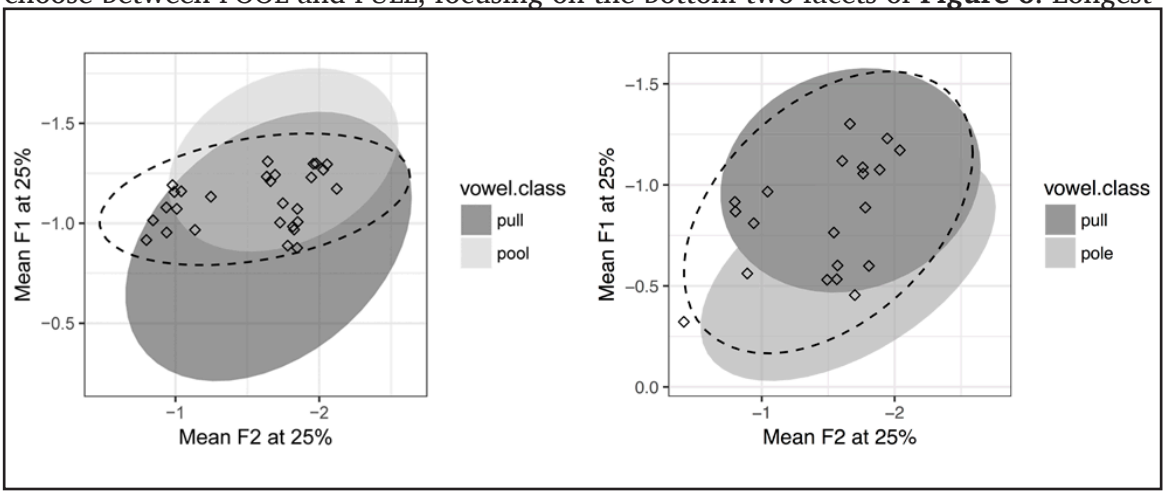

Figure 7: Merged speakers' mean productions (diamonds) compared to unmerged speakers' productions (shaded areas). Speakers who merge POOL and PULL realize the merged phoneme closer to an unmerged speaker's realization of POOL.

duration ambiguous tokens are recognized accurately around chance levels, with PULL class words being recognized on average only slightly less accurately. This is likely again due to the realization of the POOL-PULL merger, which is closer to POOL (see **Figure 7**). Participants would therefore be slightly less accurate when identifying ambiguous PULL tokens, since they would be realized closer to a distinct speaker's realization of POOL. Shorter duration ambiguous POOL tokens drop far below chance accuracy levels, while shorter duration ambiguous PULL tokens are recognized at above chance levels.

For distinct tokens, a similar pattern arises in that both word classes are accurately identified at similar levels for long tokens, then diverge for shorter duration tokens. Both PULL and POOL are recognized highly accurately when the duration of the stimulus is long, but as duration increases, accuracy for POOL tokens drops while accuracy for PULL remains mostly stable. This stability may seem contrary to predictions at first, appearing as though no duration effects exists for distinct PULL tokens. However, it is important to keep in mind that accuracy for shorter PULL tokens is expected to both drop, because shorter duration tokens are more difficult to process, and raise, because shorter tokens align with listener expectations. These two opposing influencing may yield what appears to be stability across duration categories; however, the simple fact that accuracy for PULL does not drop for shorter duration tokens is enough to suggest that a duration effect is present. It is also worth noting that, for distinct POOL-PULL tokens, accuracy for PULL remains high across duration categories and cannot realistically be expected to increase much.

3.3.1.3 The role of production in perception

The question of whether listeners integrate their own production strategies into their perceptual behaviors remains. One dimension of production that listeners might draw from in the perception task is vowel quality. That is, participants may utilize duration in perception differently depending on the degree of spectral overlap between each pair in their own speech. As **Table 3** shows, for the POOL-PULL pair, there is a main effect of Euclidean distance on accuracy; participants with greater Euclidean distance measures between POOL and PULL have generally greater accuracy ($p < .001$), though there is no effect for the POLE-PULL pair ($p = .847$). There is also a significant interaction between duration and Euclidean distance for the POOL-PULL pair, showing that participants with higher Euclidean distance measures are less influenced by duration for POOL ($p < .001$), though this interaction is again not significant for the POLE-PULL pair ($p = .119$). Finally, there is a significant three-way interaction between duration, Euclidean distance, and word class for the POOL-PULL pair ($p < .001$) but not for the POLE-PULL pair ($p = .068$). Not only are none of these three factors significant predictors in the POLE-PULL model, but the estimates are not in the expected direction (i.e., they are all in the opposite direction of the corresponding predictors in the POOL-PULL model). Only for the POOL-PULL pair, then, are unmerged speakers overall more accurate and less reliant on duration as a perceptual cue for both word classes. This pattern is summed up in **Figure 8**, which plots the correlations between Euclidean distance and duration effect[12] for each vowel pair.

It is important to note that not all speakers conform to these overall trends. In fact, there is considerable interspeaker variation, which is apparent from **Figure 8**. Other dimensions of production may be able to account for some of this variation. For instance, it might be expected that individuals who produce greater durational distinctions will rely more heavily on duration as a perceptual cue. Each model also includes duration difference as a predictor, measured as each participant's mean tense vowel duration (POOL or POLE) minus their mean

[12] In visualizations of the extent to which duration was utilized in the duration-manipulated perception task, a duration effect was calculated by subtracting accuracy percentages of the shortest duration tokens from those of the longest duration tokens. This means that, for speakers influenced by duration, the duration effect would be positive for POLE and POOL tokens and negative for PULL tokens.

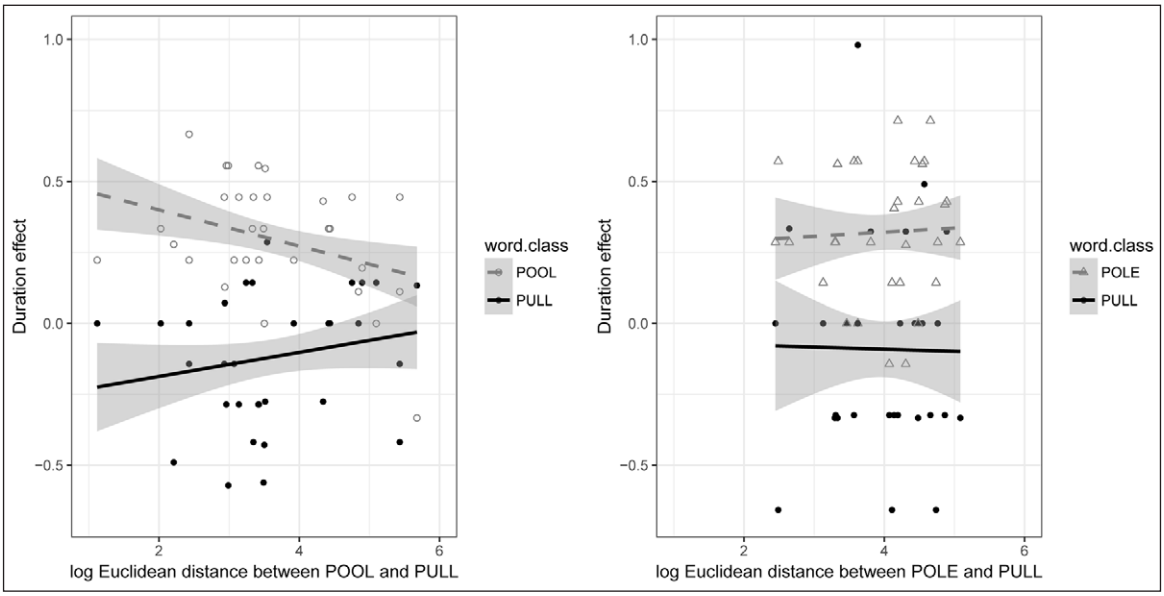

Figure 8: Youngstown participants' Euclidean distance correlation with duration effect; Euclidean distance between POOL-and PULL-class words significantly correlates with the effect of duration on perception for POOL-class words but not for PULL-class words. There is no significant correlation for Euclidean distance between POLE- and PULL-class words and duration effect.

duration for PULL. In neither model is there a main effect of duration difference, nor is there a significant two-way interaction between duration and duration difference, or three-way interaction between duration, duration difference, and word class ($p > .05$). Surprisingly, it can be concluded that utilizing duration contrastively in production does not correlate with reliance on duration in perception. **Figure 9** illustrates this lack of correlation.

There were not enough ambiguous tokens per word class per speaker to reliably determine whether accuracy correlated with speaker production differently for ambiguous or distinct tokens.

3.3.2 Burlington listeners

Analysis of Youngstown participants' performance on the perception task shows that listeners do use duration as a cue in discriminating among vowel classes. Not only are spectrally-merged Youngstown participants able to utilize durational cues in perception, but they rely on duration to a greater extent than spectrally-distinct speakers, at least for the POOL-PULL pair. From the Youngstown data alone, however, it is not easy to speculate why the community as a whole is influenced by duration. It may be the case that Youngstown listeners utilize durational cues, even when spectral cues are present, simply because durational cues enhance spectral cues and aid in identification. Alternatively, unmerged Youngstown listeners may rely on durational cues because they are more stable than the many vowel-quality realizations that exist in a community comprised of merged, unmerged, and triple-merged speakers. By comparing Youngstown's perception behavior to that of a more homogeneous community—Burlington, VT—the reasons for across-the-board reliance on durational cues may be elucidated. If Burlington listeners fail to utilize duration, we might conclude that Youngstown listeners only do so because they are consistently exposed to multiple, conflicting, spectral cues in their community.

In the aggregate, Burlington listeners appear to be following a similar pattern to that of Youngstown listeners, suggesting that they also use duration as a cue in discriminating between vowel classes. As **Tables 4** and **5** show, in both models, there is a significant main effect of duration ($p < .001$), as well as a significant interaction between duration and word class ($p < .01$). **Figure 10** illustrates this pattern, showing that accuracy for

POLE- and POOL-class words increases with duration, while accuracy for PULL-class words decreases.

Just as with Youngstown listeners, ambiguous tokens are accurately categorized at much lower rates overall (41%) than distinct tokens (80%). Both models show a significant main effect of ambiguity in this direction ($p < .001$). There is also a significant interaction between ambiguity and word class for the POOL-PULL pair, suggesting that PULL tokens show an even greater drop in accuracy when they are ambiguous ($p = .008$), though

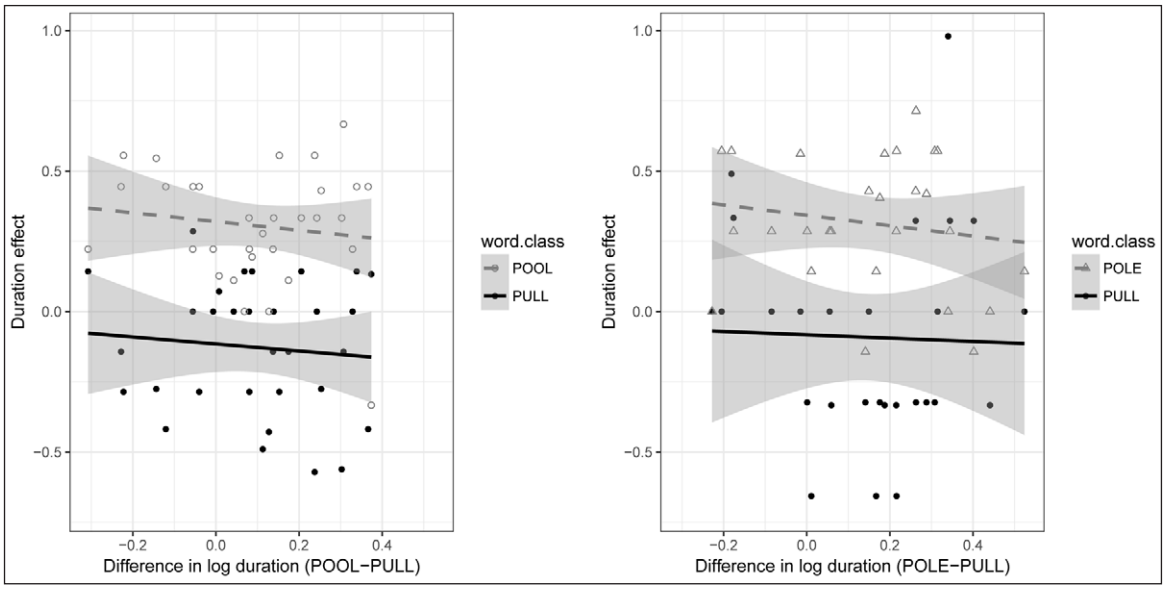

Figure 9: Lack of correlation between mean duration difference (x-axis) in production and effect of duration manipulation on perception (y-axis) for individual speakers. The word class labels to the right of each figure indicate the vowel class of the stimulus. The graph in the left facet only includes perception results for stimuli for which participants were asked to choose between POOL and PULL; the graph in the right facet presents only perception data for stimuli for which participants were asked to choose between POLE and PULL.

Table 4: POOL-PULL logistic regression model for Burlington and Youngstown listeners. Ambiguity is sum-coded while all other categorical variables are treatment coded. The reference level for word class is POOL, and the reference level for community is BURLINGTON.

	Estimate	*SE*	*z* value	Pr (>|z|)
(Intercept)	0.2849	0.1550	1.84	0.0660.
Duration	0.0037	0.0010	3.85	0.0001***
Ambiguity	−0.5972	0.0735	−8.12	<0.0001***
WordClass	0.6148	0.2463	2.50	0.0126*
Community	−0.7956	0.1725	−4.61	<0.0001***
Duration × Ambiguity	−0.0019	0.0004	−4.38	<0.0001***
Duration × WordClass	−0.0043	0.0015	−2.89	0.0039**
Ambiguity × WordClass	−0.3186	0.1192	−2.67	0.0075**
Duration × Community	0.0011	0.0011	0.99	0.3231
WordClass × Community	1.1675	0.2762	4.23	<0.0001***
Duration × Ambiguity × WordClass	0.0005	0.0007	0.77	0.4407
Duration × WordClass × Community	−0.0029	0.0016	−1.80	0.0726

Table 5: POLE-PULL logistic regression model for Burlington and Youngstown listeners. Ambiguity is sum-coded while all other categorical variables are treatment coded. The reference level for word class is POLE, and the reference level for community is BURLINGTON.

	Estimate	*SE*	*z* value	Pr (>|*z*|)
(Intercept)	−0.5716	0.1701	−3.36	0.0008***
Duration	0.0044	0.0010	4.49	<0.0001***
Ambiguity	−0.5647	0.0768	−7.35	<0.0001***
WordClass	2.6519	0.3870	6.85	<0.0001***
Community	0.4373	0.1893	2.31	0.0209*
Duration × Ambiguity	−0.0009	0.0005	−1.91	0.0561
Duration × WordClass	−0.0115	0.0020	−5.74	<0.0001***
Ambiguity × WordClass	0.5977	0.1517	3.94	0.0001***
Duration × Community	0.0014	0.0011	1.26	0.2063
WordClass × Community	−1.7174	0.4140	−4.15	<0.0001***
Duration × Ambiguity x WordClass	0.0021	0.0009	2.43	0.0152*
Duration × WordClass × Community	0.0048	0.0022	2.20	0.0277*

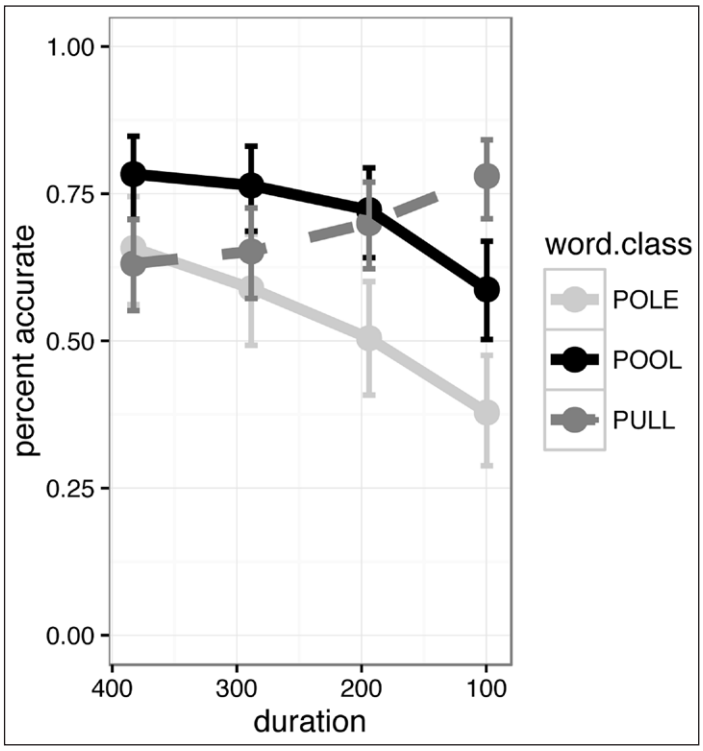

Figure 10: Burlington participants' aggregate accuracy for each duration category (95% confidence intervals).

this trend does not extend to PULL for the POLE-PULL pair.[13] Compared to Youngstown listeners, there is a somewhat more conclusive effect of ambiguity on the extent to which duration is utilized for Burlington listeners. Despite predictions that listeners would be

[13] In fact, the effect for the POLE-PULL pair is in the opposite direction.

more reliant on durational cues for ambiguous tokens, Burlington speakers actually show the opposite pattern. For the POOL-PULL pair, there is an interaction between duration and ambiguity, suggesting that listeners are less influenced by duration for ambiguous POOL tokens ($p < .001$). Lack of a three-way interaction between duration, ambiguity, and word class suggests that this trend does not extend to PULL tokens ($p = .44$). This is unsurprising because accuracy for distinct PULL is close to 100% for all durations (**Figure 11**, lower right facet), so it is not likely that there could be even less of an influence of duration for ambiguous tokens. For POLE-PULL tokens, there is perhaps a moderate interaction between ambiguity and duration, such that listeners are also less influenced by duration for ambiguous POLE tokens ($p = .056$). There is also a significant three-way interaction for the POLE-PULL pair between duration, ambiguity, and word class, suggesting that this trend extends to PULL tokens as well ($p = .015$). Burlington speakers appear to be most influenced by duration when durational cues are combined with—and enhance—already strong spectral cues. This trend will be further explored in the discussion section.

When accuracy for each word class is faceted by ambiguity and response choices given, many of the patterns are quite similar to those of Youngstown listeners. For POLE-PULL tokens (top two facets of **Figure 11**), for instance, ambiguous PULL is always more accurate than ambiguous POLE, and distinct POLE-PULL tokens show a crossover pattern such that

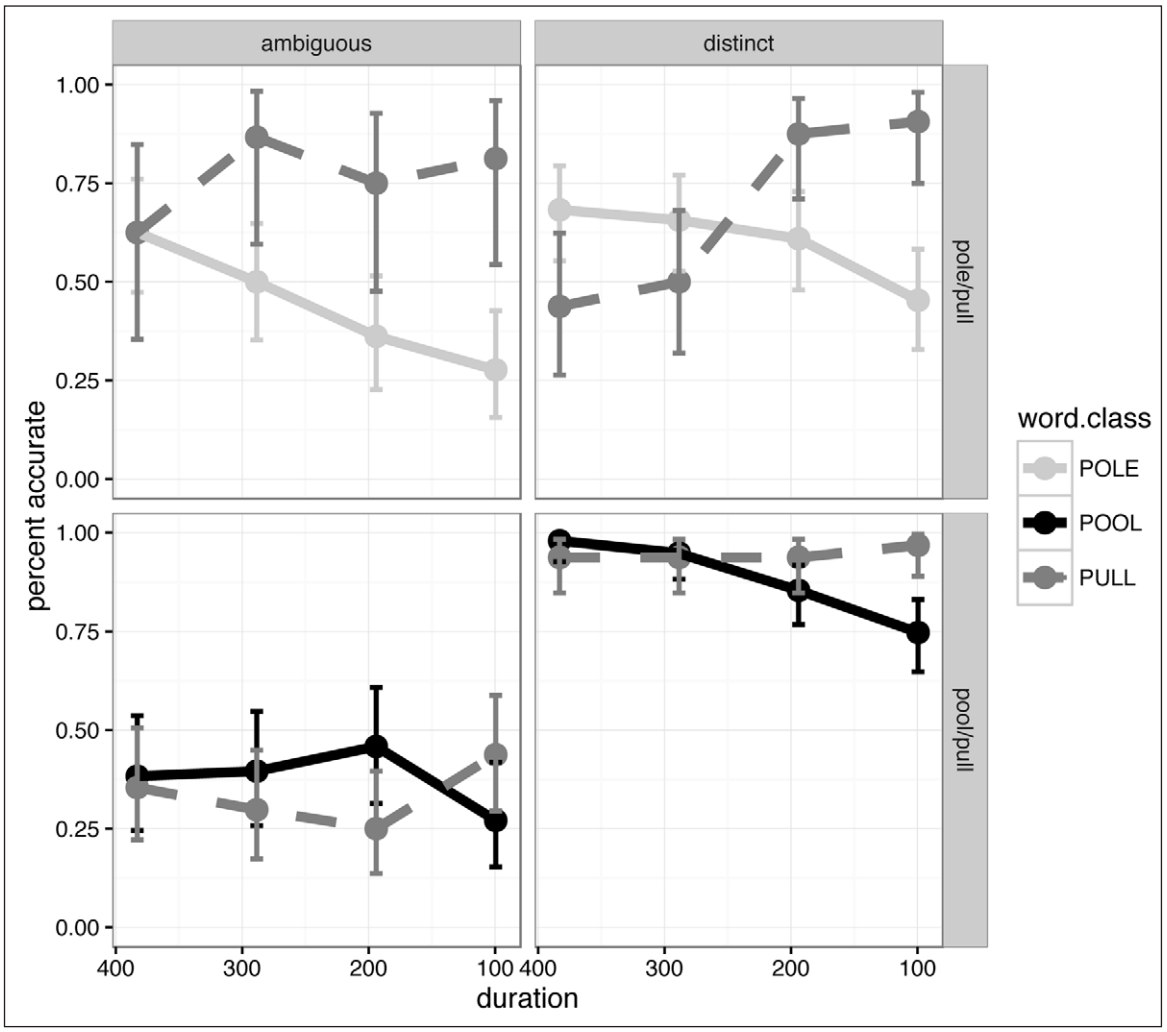

Figure 11: Burlington participants' accuracy rates by ambiguity of stimuli and response options given (95% confidence intervals).

accuracy for PULL is lower than that of POLE for longer duration but higher for shorter durations. For the POOL-PULL pair (bottom two facets), there are also many similarities for distinct tokens; accuracy for PULL is overall high and does not vary much across duration categories, while accuracy for POOL starts off equally high for the longest duration tokens but steadily decreases as duration becomes shorter.

3.3.2.1 POOL vs. PULL stimuli

There are, however, some notable descriptive differences between Youngstown and Burlington listeners. Perhaps the clearest difference is in the accuracy rates for ambiguous POOL-PULL tokens in the lower left quadrant (**Figure 11**). Youngstown participants show a clear duration effect for ambiguous tokens, while Burlington listeners' accuracy actually correlates with duration in the opposite direction of what would be expected up until accuracy for the shortest duration tokens reverses this pattern. Even for distinct POOL-PULL tokens, it appears as though Burlington listeners are less reliant on duration, evidenced by the shallower slope for POOL tokens. However, this trend of Burlington listeners appearing to be less influenced by duration for POOL-PULL tokens is not statistically significant ($p > .05$).

3.3.2.2 POLE vs. PULL stimuli

Though the patterns between Youngstown and Burlington listeners for POLE-PULL stimuli are quite similar, there are also some differences worth noting. For instance, there appears to be a larger jump in accuracy for distinct PULL tokens between the 200 and 300 marks. This crossover pattern is much subtler for Youngstown listeners and occurs between the 200 and 100 marks. In fact, there is a significant interaction between duration, word class, and community ($p = .028$), suggesting that Burlington listeners are actually *more* influenced by duration for PULL tokens when choosing between POLE and PULL than Youngstown speakers are. Though this result is initially surprising, it makes sense in light of Burlington production data.

As **Figure 12** shows, POLE and PULL are much closer in the vowel space (i.e., smaller Euclidean distance) for Burlington participants than POOL and PULL are (Est. = 79.36, $p = .002$). The close proximity of Burlington participants' phonetic targets for POLE and PULL likely affects their discriminatory abilities as well. As **Figure 13** shows, on average, the duration effect for the POOL/PULL pair is much closer to zero, and 95% confidence intervals overlap with zero for PULL. Contrastively, for the POLE-PULL pair, the differences in duration effect are much greater. This is largely due to the fact that listeners rely extensively on duration for PULL. This difference in production may also explain why Burlington listeners are more accurate for POOL in general (**Figure 10**), while Youngstown listeners are more accurate for POLE (**Figure 5**).

Comparing Youngstown listeners to Burlington listeners allows us to conclude several things about the role of the community in speech perception. First, listeners have a general tendency to be influenced by duration in categorization tasks regardless of community patterns. This does not mean that community patterns do not play a role in perceptual tendencies. On the contrary, Burlington listeners seem to show a greater duration effect for the POLE-PULL pair because it is realized closer together in the vowel space than the POOL-PULL pair. Similarly, in Youngstown, production and perception fail to align in numerous dimensions; most notably, participants who produce greater durational distinctions do not appear to rely more heavily on duration in perception. In cases of production-perception asymmetries, it must be the case that perceptual patterns are influenced by something other than an individual's production—likely, linguistic experience garnered from communication with other members of the community. Still, duration seems to be utilized most when vowel quality cues are unavailable in one's own speech (as is the case

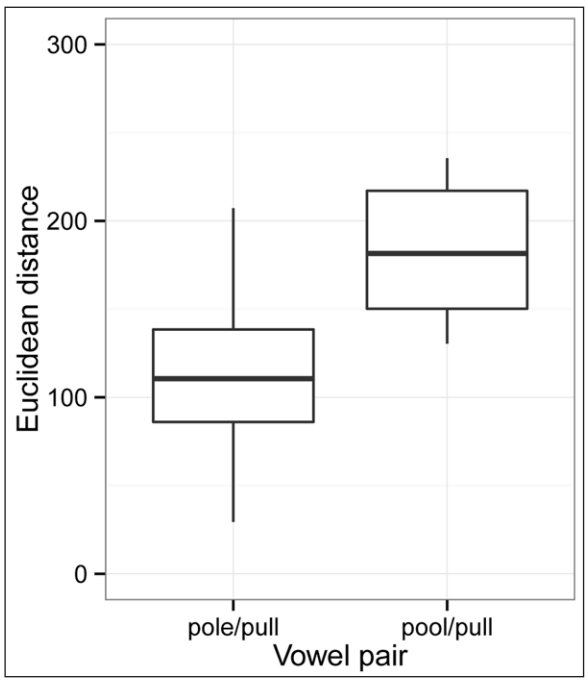

Figure 12: Burlington speakers' Euclidean distance between each vowel pair.

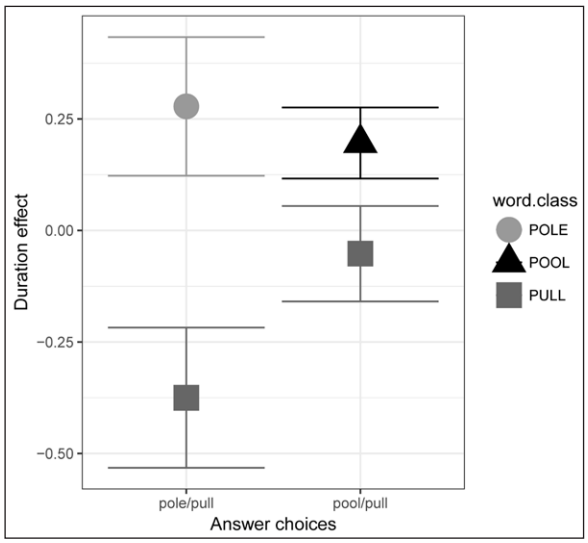

Figure 13: Burlington participants' duration effect for each vowel class.

with Youngstown POOL-PULL) or are diminished in the community (as is the case with Burlington POLE-PULL).

4 General discussion

4.1 Duration in production

As a group, spectrally-merged speakers in the Youngstown community do not make *larger* durational distinctions in production than unmerged speakers for either merger pair, evidence perhaps against a cue-trading production relationship between temporal and spectral cues. In other words, the durational distinctions made between vowel classes in the Youngstown community appear to be *maintained* by the spectrally-merged speakers who produce them (in line with Fridland et al., 2014), rather than enhanced or innovated

by them (e.g., Labov & Baranowski, 2006). The following remains to be answered: How do spectrally-merged speakers go about maintaining durational contrasts and what is the phonemic status of such contrasts?

One possibility is that spectrally-merged speakers who utilize duration in production are actually phonemically merged and learn durational distinctions through exposure to unmerged speech in their community. Phonetically rich models of speech perception (e.g., Goldinger, 1996; Johnson, 1997; Pierrehumbert, 2001) may be well suited to explain such behavior. For instance, recent studies have provided evidence that homophones can systematically differ in production (Hay & Bresnan, 2006; Gahl, 2008; Drager, 2011), and these differences can be utilized in perception (Drager, 2011). If phonemically merged speakers store phonetically rich exemplars of unmerged speech, they may be able to assign the correct durational distributions to lexical items based on this stored knowledge, even if they have only one underlying category. It no doubt helps that this is a conditioned merger and the phonemic distinction still persists in non-pre-lateral environments. Even if vowel quality differences present in unmerged speech are not mimicked in merged speech production, this does not mean that durational differences cannot be. Durational distributions might be easier to acquire than subtle vowel quality distinctions that do not exist in a speaker's phonology or whose phonetic targets a speaker has not had practice reaching in particular phonological environments (i.e., before /l/). However, why spectrally-merged speakers would more readily pick up and utilize durational distinctions for only one merger pair is not clear.

Another possibility is that durational cues were never lost in production for some speakers. Rather, spectral merger is just that—a merger of vowel quality and nothing more. Similar to what is proposed in coarticulatory models of sound change (e.g., Beddor, 2009; Beddor et al., 2002), it seems likely that durational cues were present due to articulatory constraints in producing tense and lax vowels, and when vowel quality cues disappeared, durational cues remained. This explanation would imply that some varieties of American English might be described as having phonemic vowel length, at least in some prosodic environments.

4.2 Production-perception relationships

Neither explanation proposed for the mismatch in use of spectral and durational cues in production, however, can easily account for individual production-perception mismatches regarding duration. Recall that the there was no correlation between use of duration in production in Experiment 1 and reliance on duration in perception in Experiment 2. If spectrally-merged speakers who utilize duration in production simply never lost the durational contrast, we would expect to observe durational distinctions in production as well as perception. Similarly, if speakers were storing lexically-specific, phonetically rich information of these vowel classes from exemplars heard in the community, it would also be expected that duration usage in production and perception would align, though this is not the case.

One explanation could be that durational cues present in community production patterns have influenced the cue weighting strategies of spectrally-merged listeners but have not altered their individual production patterns. It is not apparent that a general cue weighting strategy in perception necessitates the same ranking in an individual's own speech production. For instance, Shultz et al. (2012) and Francis et al. (2008) found that cue weighting for the same phonetic contrast was different in production and perception. This explanation hinges on the assumption that durational cues are weighted higher in

perception because durational cues are somehow easier to access than spectral cues, especially for non-native[14] contrasts (e.g., Cebrian, 2006).

Another explanation may stem from the fact that the mergers of interest in this study are conditioned mergers, occurring only in pre-lateral environments. Warner et al. (2004), for instance, found that for Dutch final devoiced consonants, participants were able to perceive a duration distinction that does not systematically exist in production by extending cues present in another environment (intervocalically) to the neutralized environment (word-finally). Participants in the present study could be doing the same, but with duration rather than vowel quality. This would require either reliance on orthographic patterns, which are not always predictable, or on some knowledge of vowel class membership, even in pre-lateral environments. An important point to underscore here, which makes such an explanation possible, is that speakers who fail to utilize both spectral and durational distinctions in production are not necessarily phonemically merged. On the contrary, it is possible that some individuals have two underlying phonological categories even if their production is entirely (both spectrally and otherwise) merged (Hay et al., 2010, 2013).

It is also likely that the explanation for spectrally-merged speakers' use of duration differentially in production or perception varies depending on the merger pair, particularly since the production patterns are different between the vowel pairs. There is some evidence that the POLE-PULL merger has never lost its durational contrast and maintains a phonemic length distinction in Youngstown English, as spectrally-merged speakers utilize duration in production just as much as unmerged speakers. Speakers spectrally merged between POOL and PULL, on the other hand, might be phonemically merged but able to utilize durational contrasts in perception because it is an easy cue to access.

4.3 Linguistic experience

When individuals fail to show production-perception symmetries, it must be the case that their perceptual patterns are influenced by something besides their own production, such as input from other members of the speech community. The role of the speech community in production-perception mismatches has been well-documented. For instance, NEAR-SQUARE merged NZE speakers have been shown to have some knowledge of unmerged phonologies (Hay et al., 2006; Thomas & Hay, 2005; Warren & Hay, 2006). Likewise, in Youngstown, merged speakers must have some knowledge of unmerged patterns in order to assign durational cues to the correct phonological category or lexical item.

While it is perhaps initially unexpected that merged speakers can utilize durational cues in perception, it is equally interesting that unmerged speakers are also influenced by duration, particularly since vowel quality cues were also present in the perception task and did not change across duration categories. This finding supports theories of speech perception such as Auditory Enhancement Theory, which suggests that cues covary in speech and many (often redundant) cues work together to enhance relevant contrasts in a language (Kingston & Diehl, 1994, 1995; Diehl et al., 1990). If durational cues align with expectations, their combination with existing spectral cues should enhance the phonemic contrast, while if durational cues are diminished or contradictory, even if spectral cues are still present or even primary, the acoustic signal is not as informative. These results echo previous findings (e.g., Hillenbrand et al., 2000; Sawusch & Palmer, 1997) that individuals rely on a combination of cues in perception of vowels.

It was originally hypothesized that participants would rely on duration to a greater extent when vowel-quality cues were diminished (i.e., for the ambiguous stimuli presented in the

[14] I use this term in the broad sense to include merged speakers perceiving unmerged speech as perceiving a non-native contrast.

perception task), as this would mirror previous findings that duration plays a larger role in perception of vowels that are more likely to be confused as other vowels, such as those located more centrally in the vowel space (e.g., Ainsworth, 1972) and would support the notion of a cue-trading relationship between spectral and temporal cues for these vowel pairs. However, the results of this experiment do not provide any evidence that durational cues are utilized to a greater extent for ambiguous stimuli. In fact, the opposite is true for Burlington speakers. Though it is not the case that listeners rely on duration to a greater extent when spectral cues are diminished *in that particular instance*, the uninformativeness of spectral cues does seem to mediate reliance on duration in the long-term. That is, duration is utilized more for vowel pairs that (1) are realized closer together on average in the community (2) listeners fail to produce a spectral distinction for in their own speech. This suggests that linguistic experience greatly informs reliance on duration. It is precisely the role of linguistic experience—particularly when that experience involves exposure to multiple community patterns—that I suggest in the following section contributes to misunderstanding of near-merger.

4.4 The role of duration in near-merger

Exposure to multiple community patterns means exposure to multiple cues signaling the same contrast—some contradictory. Contradictory cues have been shown to influence speech perception negatively. Fitch et al. (1980), for instance, found that discrimination of phonemes was poorer when two cues conflicted than when only a single cue was present. Sato et al. (2012) more recently found similar results with Japanese infants. We know that Youngstown speakers must produce contradictory cues because of the results of Experiment 1. Some speakers do not use duration contrastively in the expected direction, some do not use vowel quality cues informatively, and some use neither or both. Further, spectrally-merged speakers in the community may use a wide range of spectral cues to signal the merged vowel class. For these reasons, it is clear that the input listeners are receiving from the community is bound to be contradictory, at least some of the time.

I propose that contradictory cues, in combination with a misweighting of primary and secondary cues, would likely produce such results. Recall that the defining feature of near-merger is that speakers cannot identify distinctions, even if they produce them. Also recall that the standard for being able to identify distinctions is often perfect accuracy (e.g., Bowie, 2000; Labov, 1994; Labov et al., 1972). A likely explanation is not that these listeners are entirely unable to perceive the relevant contrast, but that inconsistent cues or alternative cue weighting strategies may not allow for consistent disambiguation. There are a number of possibilities for what this might look like.

If spectral cues and durational cues are contradictory in a given acoustic signal, an unmerged listener should have no problem discriminating between vowel classes if he or she weighs spectral cues higher. After all, stimuli for commutation tasks are generally chosen by their spectrally contrastive properties. However, if the acoustic signal contains contradictory cues and the listeners misweigh cues, as may be the case for non-native contrasts (e.g., Iverson et al., 2003), or relies only on secondary cues, accuracy would certainly be diminished. Durational cues may or may not be reliably present in the signal since stimuli for such tasks are often chosen for their spectral distinction alone and taken from isolated speech which may not contain useful durational information. When durational cues are present, listeners may be more accurate than when they are not, producing the above-chance yet below-100% results we often see in commutation tasks. Another possibility is that this pattern results from inconsistent weighting of primary

and secondary cues, which may stem from inconsistent input from the community. It is no surprise that near-merger is often found in communities with multiple patterns of merger and distinction, and therefore multiple realizations for the same vowel classes and inconsistent use of primary and secondary cues in signaling contrasts.

Further, it is perhaps also unsurprising that alternative distinctions often play a role in near-merger. Not only are the distinctions in dimensions such as duration or phonation relatively subtle, but they also involve differences that are not typically phonologically contrastive in English. It is well documented that it is difficult for L2 learners to produce and perceive contrasts in dimensions that are not contrastive in their L1 (e.g., Japanese speakers' difficulty with English /l/ and /ɹ/ contrast [Iverson et al., 2003]). It is likely similarly difficult for speakers to judge sounds as contrastive in their L1 when the sounds differ only along a dimension that is not generally contrastive elsewhere in the language. For instance, if a speaker has no phonemic durational contrasts in his or her grammar, besides a single durational contrast between two otherwise merged phonemes, and is then asked whether two pairs that differ only in duration are pronounced the same, the answer will likely be "yes." After all, other vowels in that speaker's inventory may be realized with various lengths and still indicate a single phoneme. In other words, when asking whether vowel pairs are 'different' and they differ only in a dimension that is not considered 'different' enough to distinguish any other vowels, a mismatch between production and judgment is not surprising. This does not mean that perceptual discrimination is not possible (see for instance Hisagi et al., 2010, on English speakers' ability to discriminate the Japanese vowel length contrast at above-chance levels) though it may be less accurate and not align with perceptual judgments.

Overall, the results of the combined production and perception data from these two speech communities add to existing evidence that secondary cues play an important part in speech production and perception.

5 Conclusion

This paper has presented evidence that spectrally-merged speakers can utilize duration to differentiate vowel classes in production, as well as recognize durational cues and associate them with the appropriate vowel classes in perception. However, duration use in production does not directly inform reliance of duration in perception, and individual differences exist in regard to whether and to what extent duration is used in both production and perception, regardless of spectral patterns. Overall, Youngstown participants who merge POOL and PULL rely more heavily on duration in perception than unmerged participants. Additionally, the POLE-PULL pair, which for Burlington speakers is much closer in the vowel space than the POOL-PULL pair, exhibits the clearest duration effects for Burlington speakers. To summarize, both merged and unmerged speakers rely on durational cues in perception to some extent; however, speakers with a greater degree of spectral overlap in production rely on duration to a greater extent. It is not just an individual's phonological system, but also his or her linguistic experience that informs patterns of production and perception. Finally, alternative cues such as duration may not only play a role in contrasting vowel classes but can also contribute to the production-perception asymmetries that define linguistic phenomena such as near-merger and should be considered in studies of phonemic contrasts.

Additional File

The additional file for this article can be found as follows:

- **Appendix.** Token counts by speaker. DOI: https://doi.org/10.5334/labphon.54.s1

Acknowledgements

Without the cooperation of the 57 participants from Youngstown, OH and Burlington, VT, this work would not have been possible. I am also grateful to Maeve Eberhardt and Toni Cook, who helped me find participants in Burlington, and Daniel Szeredi, who set up the online version of the perception experiment. This work was greatly improved by feedback from Meredith Tamminga, Erik Thomas, Robin Dodsworth, Jeff Mielke, Walt Wolfram, and the audiences of NWAV 44 and SVALP. Finally, I would like to thank associate editor Lisa Davidson and three anonymous reviewers for their thorough comments and extremely helpful suggestions.

Competing Interests

The author has no competing interests to declare.

References

Ainsworth, W. A. 1972. Duration as a cue in the recognition of synthetic vowels. *The Journal of the Acoustical Society of America*, *99*, 2350–2357. DOI: https://doi.org/10.1121/1.1912889

Arnold, L. R. 2015. Multiple mergers: Production and perception of three pre-/l/ mergers in youngstown, ohio. University of Pennsylvania Working Papers in Linguistics, *21*(2), Article 2. Retrieved from: http://repository.upenn.edu/pwpl/vol21/iss2/2

Becker, M., & Levine, J. 2014. *Experigen–an online experiment platform*. Retrieved from: http://becker.phonologist.org/experigen.

Beddor, P. S. 2009. A coarticulatory path to sound change. *Language*, *85*(4).

Beddor, P. S., Harnsberger, J. D., & Lindemann, S. 2002. Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates. *Journal of Phonetics*, *30*, 591–627. DOI: https://doi.org/10.1006/jpho.2002.0177

Bennet, D. 1968. Spectral form and duration as cues in the recognition of English and German vowels. *Language and Speech*, *11*(2), 65–85. DOI: https://doi.org/10.1177/002383096801100201

Benson, E. J., Fox, M. J., & Balkman, J. 2011. The bag that Scott bought: The low vowels in northwest Wisconsin. *American Speech*, *86*(3), 271–311. DOI: https://doi.org/10.1215/00031283-1503910

Boersma, P., & Weenick, D. 2016. *Praat: Doing phonetics by computer. Version 6.0.19*. Computer Program. Retrieved from: http://www.praat.org/

Bohn, O. 1995. Speech perception and linguistic experience: Issues in cross-language research. In: Strange, W. (ed.), 279–304. Timonium, MD: York Press.

Bowie, D. 2000. *The effect of geographic mobility on the retention of a local dialect*. (Doctoral dissertation), Philadelphia, PA: University of Pennsylvania. Retrieved from: http://repository.upenn.edu/dissertations/AAI9965448

Cebrian, J. 2006. Experience and the use of non-native duration in L2 vowel categorization. *Journal of Phonetics, 34*, 372–387. DOI: https://doi.org/10.1016/j.wocn.2005.08.003

Chen, M. 1970. Vowel length variation as a function of the voicing of the consonant environment. *Phonetica*, *22*, 129–159. DOI: https://doi.org/10.1159/000259312

Clopper, C. G., Pisoni, D. B., & de Jong, K. 2005. Acoustic characteristics of the vowel systems of six regional varieties of American English. *Journal of the Acoustical Society of America*, *118*, 1661–1676. DOI: https://doi.org/10.1121/1.2000774

De Jong, N. H., & Wempe, T. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390. DOI: https://doi.org/10.3758/BRM.41.2.385

Diehl, R. L., Kluender, K. R., & Walsh, M. A. 1990. Advances in speech, hearing and language processing. In: Ainsworth, W. A. (ed.), 243–268. London: JAI Press.

Di Paolo, M., & Faber, A. 1990. Phonation differences and the phonetic content of the tense-lax contrast in Utah English. *Language Variation and Change*, *2*, 155–204. DOI: https://doi.org/10.1017/S0954394500000326

Drager, K. K. 2011. Sociophonetic variation and the lemma. *Journal of Phonetics*, *39*(4), 694–707. DOI: https://doi.org/10.1016/j.wocn.2011.08.005

Escudero, P. 2000. *Developmental patterns in the adult L2 acquisition of new contrasts: The acoustic cue weighting in the perception of Scottish tense/lax vowels by Spanish speakers* (Unpublished master's thesis). Edindurgh: University of Edinburgh.

Faber, A., & Di Paolo, M. 1995. The discriminability of nearly merged sounds. *Language Variation and Change*, *7*, 35–78. DOI: https://doi.org/10.1017/S0954394500000892

Fitch, H. L., Hawles, T., Erickson, D. M., & Liberman, A. M. 1980. Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception and Psychophysics*, *27*(4), 343–350. DOI: https://doi.org/10.3758/BF03206123

Francis, A. L., Kaganovich, N., & Driscoll-Huber, C. 2008. Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *The Journal of the Acoustical Society of America*, *124*(2), 1234–1251. DOI: https://doi.org/10.1121/1.2945161

Fridland, V., Kendall, T., & Farrington, C. 2014. Durational and spectral differences in American English vowels: Dialect variation within and across regions. *The Journal of the Acoustical Society of America*, *136*(1), 341–349. DOI: https://doi.org/10.1121/1.4883599

Gahl, S. 2008. *Time* and *thyme* are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*, *84*(3), 474–496. DOI: https://doi.org/10.1353/lan.0.0035

Goldinger, S. D. 1996. Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22*(5), 1166–1183. DOI: https://doi.org/10.1037/0278-7393.22.5.1166

Hay, J., & Bresnan, J. 2006. Spoken syntax: The phonetics of giving a hand in New Zealand English. *The Linguistic Review*, *23*, 321–349. DOI: https://doi.org/10.1515/TLR.2006.013

Hay, J., Drager, K., & Thomas, B. 2013. Using nonsense words to investigate vowel merger. *English Language and Linguistics*, *17*(2), 241–269. DOI: https://doi.org/10.1017/S1360674313000026

Hay, J., Drager, K., & Warren, P. 2010. Short-term exposure to one dialect affects processing of another. *Language and Speech*, *53*(4), 447–471. DOI: https://doi.org/10.1177/0023830910372489

Hay, J., Warren, P., & Drager, K. 2006. Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, *34*(4), 458–484. DOI: https://doi.org/10.1016/j.wocn.2005.10.001

Herold, R. 1990. Mechanisms of merger: The implementation and distribution of the low back merger in eastern Pennsylvania. (Doctoral dissertation). Philadelphia, PA: University of Pennsylvania. Retrieved from: http://repository.upenn.edu/dissertations/AAI9026574.

Hillenbrand, J. M., Clark, M. J., & Houde, R. A. 2000. Some effects of duration on vowel recognition. *The Journal of the Acoustical Society of America*, *108*(6), 3013–3022. DOI: https://doi.org/10.1121/1.1323463

Hisagi, M., Shafer, V. L., Strange, W., & Sussman, E. S. 2010. Perception of a Japanese vowel length contrast by Japanese and American English listeners: Behavioral and

electrophysiological measures. *Brain Research*, *1360*, 89–105. DOI: https://doi.org/10.1016/j.brainres.2010.08.092

Howell, P. 1993. Cue trading in the production and perception of vowel stress. *The Journal of the Acoustical Society of America*, *94*(4), 2063–2073. DOI: https://doi.org/10.1121/1.407479

Irons, T. L. 2007. On the status of low back vowels in Kentucky English: More evidence of merger. *Language Variation and Change*, *19*(2), 137–180. DOI: https://doi.org/10.1017/S0954394507070056

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. 2003. A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47–B57. Retrieved from: http://www.sciencedirect.com/science/article/pii/S0010027702001981. DOI: https://doi.org/10.1016/S0010-0277(02)00198-1

Jacewicz, E., Fox, R. A., & Salmons, J. 2007. Vowel duration in three American English dialects. *American Speech*, *82*(4), 367–385. DOI: https://doi.org/10.1215/00031283-2007-024

Jessen, M. 1998. *Phonetics and phonology of tense and lax obstruents in German*. Amsterdam; Philadelphia: John Benjamins Publishing Co. Retrieved from: https://hdl.handle.net/2027/mdp.39015043129074.

Johnson, K. 1997. Speech perception without speaker normalization: An exemplar model. In: Johnson, K., & Mullennix, J. W. (eds.), *Talker variability in speech processing*, 145–165. San Diego: Academic Press.

Kingston, J., & Diehl, R. L. 1994. Phonetic knowledge. *Language*, *70*(3), 419–454. DOI: https://doi.org/10.1353/lan.1994.0023

Kingston, J., & Diehl, R. L. 1995. Intermediate properties in the perception of distinctive feature values. In: Connell, B., & Arvaniti, A. (eds.), *Phonology and phonetic evidence: Papers in laboratory phonology*, *IV*. Cambridge, UK: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511554315.002

Kohler, K. J. 1979. Dimensions in the perception of fortis and lenis plosives. *Phonetica*, *36*, 332–343. DOI: https://doi.org/10.1159/000259970

Labov, W. 1994. *Principles of linguistic change, Vol. 1. Internal factors.* Oxford: Blackwell.

Labov, W., Ash, S., & Boberg, C. 2006. *The atlas of North American English: Phonetics, phonology and sound change*. Berlin: Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110167467

Labov, W., & Baranowski, M. 2006. 50 msec. *Language Variation and Change*, *18*, 1–18. DOI: https://doi.org/10.1017/S095439450606011X

Labov, W., Karen, M., & Miller, C. 1991. Near-mergers and the suspension of phonemic contrast. *Language Variation and Change*, *3*, 33–74. DOI: https://doi.org/10.1017/S0954394500000442

Labov, W., Yaeger, M., & Steiner, R. 1972. *A quantitative study of sound change in progress*, *1*, U. S. Regional Survey, 1972.

Langstrof, C. 2009. On the role of vowel duration in the New Zealand English front vowel shift. *Language Variation and Change*, *21*, 437–453. DOI: https://doi.org/10.1017/S0954394509990159

Lobanov, B. 1971. Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America*, *68*, 1636–1642. DOI: https://doi.org/10.1121/1.1912396

Maguire, W., Clark, L., & Watson, K. 2013. Introduction: What are mergers and can they be reversed? *English Language and Linguistics*, *17*(02), 229–239. Jun. DOI: https://doi.org/10.1017/S1360674313000014

Majors, T. 2005. Low back vowel merger in Missouri speech: Acoustic description and explanation. *American Speech, 80*, 165–179. DOI: https://doi.org/10.1215/00031283-80-2-165

Milroy, J., & Harris, J. 1980. When is a merger not a merger? The meat/mate problem in a present-day English vernacular. *English World Wide, 1*, 199–210. DOI: https://doi.org/10.1075/eww.1.2.03mil

Nycz, J. 2013. New contrast acquisition: Methodological issues and theoretical implications. *English Language and Linguistics, 17*(2), 325–357. DOI: https://doi.org/10.1017/S1360674313000051

Peterson, G. E., & Lehiste, I. 1960. Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America, 32*, 693–703. DOI: https://doi.org/10.1121/1.1908183

Pierrehumbert, J. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In: Bybee, J., & Hopper, P. J. (eds.), *Frequency effects and emergent grammar*, 137–158. Amsterdam: John Benjamins Publishing Co. DOI: https://doi.org/10.1075/tsl.45.08pie

Raphael, L. J. 1972. Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *The Journal of the Acoustical Society of America, 51*, 1296–1303. DOI: https://doi.org/10.1121/1.1912974

Repp, B. 1983. Trading relations among acoustic cues in speech perception are largely a result of phonetic categorization. *Speech Communication, 2*, 341–361. DOI: https://doi.org/10.1016/0167-6393(83)90050-X

Sato, Y., Kato, M., & Mazuka, R. 2012. Development of single/geminate obstruent discrimination by Japanese infants: Early integration of durational and nondurational cues. *Developmental Psychology, 48*(1), 18–34. DOI: https://doi.org/10.1037/a0025528

Sawusch, J. R., & Palmer, N. J. (1997). The role of vowel duration in the perception of /ɛ/ and /æ/. *The Journal of the Acoustical Society of America, 101*(5), 3111–3112. DOI: https://doi.org/10.1121/1.418887

Shultz, A. A., Francis, A. L., & Llanos, F. 2012. Differential cue weighting in perception and production of consonant voicing. *The Journal of the Acoustical Society of America, 132*, EL95–EL101. DOI: https://doi.org/10.1121/1.4736711

Tauberer, J., & Evanini, K. 2009. Intrinsic vowel duration and the post-vocalic voicing effect: Some evidence from dialects of North American English. In: *Intrinsic vowel duration and the postvocalic voicing effect: Some evidence from dialects of North American English*.

Thomas, B., & Hay, J. 2005. A pleasant malady: The Ellen/Allan merger in New Zealand English. *Te Reo, 48*, 69–93.

Warner, N., Jongman, A., Sereno, J., & Kemps, R. 2004. Incomplete neutralization and other subphonemic durational differences in production and perception: Evidence from Dutch. *Journal of Phonetics, 32*(2), 251–276. DOI: https://doi.org/10.1016/S0095-4470(03)00032-9

Warren, P., & Hay, J. 2006. Using sound change to explore the mental lexicon. In: Fletcher-Flinn, C., & Haberman,, G. (eds.), *Cognition, language and development: Perspectives from New Zealand*, 101–121. Bowen Hills: Australian Academic Press.

Wassink, A. B. 2006. A geometric representation of spectral and temporal vowel features: Quantification of vowel overlap in three linguistic varieties. *The Journal of the Acoustical Society of America, 119*(4), 2334–2350. DOI: https://doi.org/10.1121/1.2168414

Watson, C. L., & Harrington, J. 1990. Acoustic evidence for dynamic formant trajectories in Australian English vowels. *The Journal of the Acoustical Society of America, 106*(1), 458–468. DOI: https://doi.org/10.1121/1.427069

How to cite this article: Wade, L. 2017 The role of duration in the perception of vowel merger. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 8(1): 30, pp. 1–34, DOI: https://doi.org/10.5334/labphon.54

Submitted: 28 September 2016 **Accepted:** 04 August 2017 **Published:** 13 December 2017

]u[*Laboratory Phonology: Journal of the Association for Laboratory Phonology* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS