Laboratory Phonology
Journal of the Association for Laboratory Phonology

Zellou, G. and Pycha, A. 2018 The gradient influence of temporal extent of coarticulation on vowel and speaker perception. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 9(1): 12, pp. 1–24, DOI: https://doi.org/10.5334/labphon.118

JOURNAL ARTICLE

The gradient influence of temporal extent of coarticulation on vowel and speaker perception

Georgia Zellou[1] and Anne Pycha[2]

[1] University of California, Davis, Linguistics Department, 469 Kerr Hall, One Shields Avenue, Davis, CA 95616, US

[2] University of Wisconsin, Milwaukee, Linguistics Department, Johnston Hall Room 117, Milwaukee, WI 53211, US

Corresponding author: Georgia Zellou (gzellou@ucdavis.edu)

Coarticulation makes vowels in context acoustically different from context-free vowels. Listeners sometimes compensate by ascribing these acoustic effects to their source, but the conditions under which they do so have not yet been fully pinpointed. Ohala (1993) had suggested that acoustic effects which are temporally more distant from their source should be more susceptible to misattribution. In three experiments, we tested this hypothesis by varying the temporal extent of coda-triggered coarticulation on vowels and investigating its influence on two different perceptual behaviors: speaker-model representation and vowel-phoneme identification. Experiment 1 asked listeners to estimate speaker height based on /giC/ and /gɪC/ nonsense tokens produced by twelve female speakers. Results indicated a gradient effect: Within lax /ɪ/, greater temporal extent of coarticulation correlated with taller height judgments. Experiment 2a was similar, except that temporal extent of coarticulation in the tokens varied across a wider range of values than in Experiment 1. Results again indicated a gradient effect: Within lax /ɪ/, greater temporal extent of coarticulation correlated with taller height judgments. In Experiment 2b, listeners performed an AXB vowel-phoneme discrimination task. Results showed that greater temporal extent of coarticulation correlated with greater likelihood of listeners judging an intended /ɪ/ token to contain the vowel /ʌ/. Taken together, our results indicate that temporal extent of coarticulation affects both speaker-models and interpretation of vowel identity.

Keywords: temporal dynamics; coarticulation; perceptual compensation; speaker height perception; vowel perception; sound change

1. Introduction

The speech signal contains a great deal of variability, which is a potential source of ambiguity for the listener. Consider vowels: Although individual vowels can be classified according to their first and second formant frequencies, the value of these frequencies varies significantly from one utterance to the next. For the same vowel-phoneme, formants can vary by as much as 70% depending upon the context in which the vowel occurs, and by as much as 99% depending upon who the speaker is (Nearey, 1989). These large effects of context and speaker mean that, in order to successfully identify the intended vowel, the listener must do at least two things: perceptually compensate for contextual effects (Mann, 1980) and construct a model of the speaker (for an overview, see Johnson, 2008). Vowel identification might involve attributing the acoustic effects of adjacent sound co-production to their proper sources, such as a following consonant (Byrd, 1992). For example, in the production of English words such as *heap* /hip/ or *hip* /hɪp/, speakers typically initiate the lip closing gesture for [p] during the tongue gesture for the vowel, lowering the characteristic formants of the vowel in the vicinity of [p]. To accurately identify the vowel, the listener should adjust for this change. Constructing a

speaker model might involve, at a minimum, making an estimate of speaker height. Taller people usually have longer vocal tracts (Fitch & Giedd, 1999) and longer vocal tracts produce lower formant values than shorter vocal tracts do (Lieberman & Blumstein, 1988). Again, to accurately identify the vowel, the listener must adjust for this change. Although the parameters of these two processes differ somewhat, perceptual compensation and speaker-model construction crucially both impose the same fundamental requirement on the listener: correct attribution of acoustic variation to its proper source.

Previous research has shown that while listeners sometimes make these attributions correctly, they sometimes do not. Consider the effect of compensation on phoneme identification first. Mann (1980) showed that listeners were more likely to hear an ambiguous [d]/[g] stimulus as /d/ when the preceding sound was an /r/, compared to when it was an /l/. In this case, listeners attributed the F3-lowering in the stop transition to the preceding rhotic (rather than /g/) thus biasing their responses toward /d/. However, subsequent research has shown that compensation is not an all-or-nothing phenomenon. Beddor and Krakow (1999) showed that listeners did not perceive nasalized vowels in nasal consonant contexts (i.e., NṼN) as equivalent to oral vowels in oral contexts (CVC), suggesting that listeners only partially compensated by attributing some but not all of the acoustic effects of nasalization to the nasal consonant gesture. Thus, there are many unanswered gaps in our understanding of perceptual compensation. Although we know that speakers *can* compensate, we have not yet pinpointed the exact conditions under which they *do*, or only partially, compensate.

Previous research on speaker-models has also yielded mixed results and raises questions that are relevant to what listeners do with coarticulatory information. A number of studies have presented listeners with vowel tokens with the same identity, such has /i/, but with varying formant values. Listeners consistently reported that the token with lower formant values was produced by a taller speaker (Fitch, 1994; Ives, Smith, & Patterson, 2005; Rendall, Vokey, & Nemeth, 2007; Smith & Patterson, 2005; Smith, Patterson, Turner, Kawahara, & Irino, 2005). In these cases, listeners constructed a model of the speaker, attributing differences in formant values to speaker height. On the other hand, Barreda (2016, 2017) presented listeners with vowel tokens that differed in identity, such as /e/ versus /o/, but that were produced by the same (synthetic) vocal tract. Listeners were more likely to judge that /o/ came from a taller speaker, presumably because /o/ has an inherently lower value for F2 than /e/ does (and only a slightly different F1 value). Similar findings held for vowel pairs that differed primarily in F1, in F3, and in F1–F2 and F2–F3 combinations. In this case, listeners apparently did not construct a single, consistent model of the speaker; if they had, their height judgments for /e/ versus /o/, and the other vowel pairs, should have been equivalent. Again, there are unresolved issues in our understanding of what information in the speech signal listeners use to construct speaker-models. Although we know that speakers *can* construct consistent speaker-models, we have not yet pinpointed the exact conditions under which they *do* construct them. More specifically, when listeners fail to attribute coarticulatory information to its proper source, do they use that acoustic information in constructing a speaker-model?

How, then, should we go about identifying the precise factors that modulate listeners' attribution of acoustic variability? Ohala (1993, p. 247) offered an important but largely under-explored suggestion. He speculated that acoustic changes which are temporally more distant from their source should be more susceptible to misattribution. That is, the greater the temporal gap between the source of variation and its acoustic effect, the more difficult it should be for listeners to correctly attribute that variation to its source. Consider again *heap* /hip/ and *hip* /hɪp/. We noted that the characteristic formants of the vowel lower due to the presence of the labial coda. In addition, the temporal extent of this

coarticulatory influence can and does differ from one utterance to the next (Pycha, 2016). **Figure 1** shows three schematic examples of the American English high front lax vowel [ɪ], all with 150 millisecond durations, produced before coda [p]. In (a), the value of F2 remains steady from 0 to 125 ms, when it begins to lower due to anticipatory coarticulation with the labial coda. Thus, the temporal extent of coarticulation measures 25 ms. In (b), the temporal extent of coarticulation measures 50 ms, and in (c) it measures 75 ms. (Of course, a labial coda would also typically lower F1 values, but the steady-state F1 values of [ɪ] are already so small that additional lowering tends to be negligible, so we do not focus on it here. We are also excluding any potential influence of onset consonants, which we held constant in our experimental stimuli).

In all three vowels, the source of acoustic variation is the labial coda, and the effect is F2 lowering. If listeners properly compensate, they should attribute F2 lowering to the labial coda, not to the quality of the vowel itself or to the speaker's height. Yet across vowels,

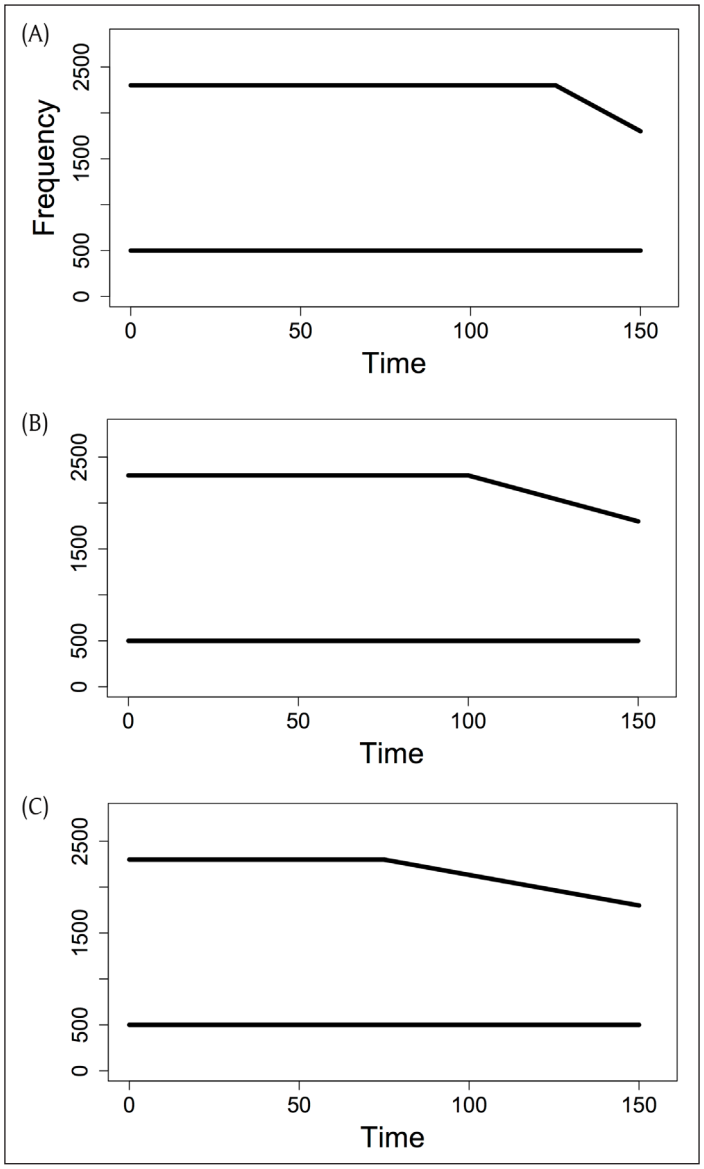

Figure 1: (a) Schematic F1 and F2 trajectories for lax [ɪ] before a labial coda, where temporal extent of F2-lowering coarticulation is 25 ms. **(b)** Schematic F1 and F2 trajectories for lax [ɪ] before a labial coda, where temporal extent of F2-lowering coarticulation is 50 ms. **(c)** Schematic F1 and F2 trajectories for lax [ɪ] before a labial coda, where temporal extent of F2-lowering coarticulation is 75 ms.

the temporal relationship between source and acoustic effect crucially differs. At one extreme, in a), the source is never more than 25 ms from the effect. At the other extreme, in c), the source can be as much as 75 ms away from its effect. The key prediction is that misattribution is unlikely for vowels with the pattern in a), somewhat more likely for vowels with the pattern in b), and much more likely in vowels with the pattern in c). In the current study, we pursue this prediction by presenting listeners with CVC tokens in which the temporal extent of coda coarticulation varies and asking them to estimate the height of the speaker or identify the quality of the vowel.

To our knowledge, only one previous study has followed up on Ohala's (1993) compelling idea. In an eye-tracking experiment, Beddor and colleagues (Beddor, McGowan, Boland, Coetzee, & Brasher, 2013) showed that, after hearing a CVNC token such as *bent* or *bend*, listeners fixated more quickly and more often on the corresponding CVNC image if coarticulatory nasalization began early in the vowel, compared to if it began late. This shows that listeners do indeed respond to stimuli differently depending upon the temporal relationship between source and acoustic effect, and therefore suggests that Ohala's (1993) speculation was on the right track. However, Beddor and colleagues did not actually ask listeners what type of vowel they perceived in these stimuli (i.e., oral versus nasal). As such, while their findings show that listeners did make use of variation, they do not reveal whether they also compensated for it. This important issue thus remains open for further investigation.

Studying the conditions under which listeners do, or do not, attribute acoustic variation to its source is an important task for the field of speech perception, because misattribution can lead to mis-interpretation of the speaker's intended utterance. For example, if a listener does not perceptually compensate for nasal coarticulation, she or he may perceive an inherently nasal vowel, even though the speaker intended an oral vowel. In support of this idea, Kawasaki (1986) showed that when the intensity of flanking nasal consonants in NṼN syllables was diminished (presumably making it more difficult to attribute the vowel nasality to its original source), listeners were more likely to identify that the vowel was nasalized, rather than oral. A similar scenario occurs whenever a listener does not build an appropriate model of the speaker. For example, Eklund and Traunmüller (1997) showed that when listeners misidentified the gender of a speaker, their error rate in vowel identification increased to 25%, compared to 9% when they correctly identified the gender.

Misattributions are of particular interest because they have the potential to trigger phonological reanalysis and sow the seeds for historical sound change (Ohala, 1993). For example, Ohala speculated that, over time, listener misattribution of nasal coarticulation gave rise to phonemically nasal vowels in French, precisely in those cases where the nasal source itself was lost (cf. French *brun* [brœ̃] 'brown'). Interestingly, we are not aware of a similar speculation for speaker-models, despite the fact that misattribution of speaker height or gender should be an equally probable source of change. Indeed, the literature on these two processes differs noticeably on this point: Previous studies of compensation have taken up Ohala's challenge and demonstrated an explicit interest in linking experimental findings to historical sound change (Beddor, 2010; Beddor, Harnsberger, & Lindemann, 2002; Beddor & Krakow, 1999; Yu, 2010; Yu & Lee, 2014; Zellou, 2017), while studies of speaker-models have neglected the issue, to our knowledge.

In fact, the literature on these two processes also differs on another important point. Previous studies of compensation have focused largely on factors that are external to the speech signal, such as individual differences in production patterns (Zellou, 2017) or personality traits (Yu, 2010). While these findings are promising, factors internal to

the speech signal—such as the temporal extent of coarticulation—also warrant attention. Meanwhile, the literature on speaker-models exhibits a more holistic approach. Some studies have demonstrated the role of factors that are external to the speech signal, such as differences in the gender stereotypicality of speaker voices (Johnson, Strand, & D'Imperio, 1999) or in the listener's beliefs about the speaker's national identity (Niedzielski, 1996), in influencing vowel identification. Other studies have demonstrated the role of factors that are internal to the speech signal, such as the formant frequency differences between /e/ and /o/ reviewed above (Barreda, 2016, 2017).

In the current study, we address these gaps and pursue Ohala's temporal gap hypothesis for perceptual compensation while treating both vowel-phoneme identification and speaker-models as potential sources for reanalysis. Specifically, we examine whether the temporal extent of acoustic variation affects the perceptual compensation that listeners perform for contextual changes, in particular those induced by a coda consonant, as realized by a) the model that listeners build of the speaker, in particular of the speaker's height, and b) correct identification of the speaker's intended vowel-phoneme.

2. Experiment 1: The effect of temporal extent of coarticulation on estimates of speaker height, using vowels with maximal differences

In Experiment 1, participants listened to nonsense gVC words that contained high, front tense or lax vowels ([git, gid, gip, gib] and [gɪt, gɪd, gɪp, gɪb]), and provided a height estimate for the speaker for each token (e.g., 5 feet, 3 inches). The crucial question was whether participants would give greater height estimates for tokens in which the vowel contained longer formant-lowering transitions, compared to those in which it had shorter ones—even when those items were produced by the same speaker. Previous work had shown that the duration of formant-lowering transitions is greater in the lax /ɪ/ compared to tense /i/ (Pycha, 2016). For this reason, we selected stimuli so as to maximize, and have non-overlapping, differences across vowel categories: Tokens with tense vowels had proportionally shorter transitions, while tokens with lax vowels had proportionally longer transitions.

2.1 Method

2.1.1 Stimuli

The perceptual stimuli were selected from the production data of Pycha (2016), in which fifteen speakers produced the nonce words [git, gid, gip, gib] and [gɪt, gɪd, gɪp, gɪb]. In that study, the onset [g] was used because it creates a nonsense word for every iC and ɪC rime; it also helps ensure that onset consonants exert similar coarticulatory effects across different tokens. We included both labial and alveolar codas because they both have formant-lowering effects on the preceding vowel (albeit to different degrees). The words were produced with voiced and voiceless codas in four different contexts: phrase-medial in unfocused position, phrase-medial in focused position, phrase-final in unfocused position, and phrase-final in focused position. These different voicing values and sentential contexts modulated the coarticulatory effects of the coda consonant on the preceding vowel, such that the data contained a wide range of values for *transition ratio*. Transition ratio is the key variable that we used to quantify the temporal extent of coarticulation, and we defined it as the duration of the F2 formant transition divided by the total duration of the vowel. For example, a transition ratio of 0.20 indicates that formant-lowering transitions occupied 20% of the vowel's total duration, while a transition ratio of 0.77 indicates that they occupied 77%. We focused on F2 because both /i/ and /ɪ/ have high steady-state values for F2, allowing us to identify the onset of F2-lowering transitions with relative ease.

By contrast, both /i/ and /ɪ/ have low steady-state values for F1 making it much more difficult to pinpoint when and where lowering occurs. The techniques for demarcating steady-state versus transitional areas of the vowel are described in Pycha (2016).

The original data set contained 960 tokens (64 from each of the 15 speakers). In order to exclude the potentially confounding effect of gender, we did not include tokens produced by the three male speakers. From the remaining set of 768 tokens, we selected 96 giC tokens (eight from each of the 12 speakers) and 96 gɪC tokens (again, eight from each of the 12 remaining speakers). We selected tokens whose values for transition ratio (i.e., temporal extent of coarticulation) differed maximally across lax versus tense conditions. That is, we chose /i/ tokens for which the transition ratio had relatively small values, and /ɪ/ tokens for which it had relatively large values. In doing so, we observed two constraints. First, in order to ensure that we included giC and gɪC tokens produced by speakers of the exact same height, we balanced the stimuli for speaker. Second, because labial codas exert a significantly stronger effect on transition ratio than alveolar codas (Pycha, 2016), we balanced the stimuli for coda place-of-articulation. Thus, we had an equal number of gi[p/b] tokens for each speaker, an equal number of gi[t/d] tokens for each speaker, and so on. To the extent possible, we also matched the spectral extent of coarticulation (defined as the F2 at steady state, divided by F2 at offset) across these conditions, such that the crucial difference between tense and lax tokens lay in the temporal extent of coarticulation, not the spectral extent (i.e., magnitude of formant-lowering coarticulation). The acoustic measurements for the 192 stimuli in Experiment 1 are displayed in **Table 1**.

These stimuli contain a number of characteristics that we will bear in mind in the interpretation of our results. Starting with the first column in **Table 1**, we see values for transition ratio, our critical variable. The mean transition ratio of tense /i/ is smaller than that of lax /ɪ/, which indicates that the temporal extent of coarticulation is greater in gɪC than in giC. This is in line with the findings of Pycha (2016) as well as our current stimulus selection procedures. The mean duration of tense /i/ is greater than that of lax /ɪ/, aligning with previous descriptions (e.g., Peterson & Lehiste, 1960). The mean F1 steady-state of tense /i/ is smaller than that of lax /ɪ/, also in agreement with previous findings (e.g., Stevens & House, 1963). The mean F1 ratio for tense /i/ is very close to 1, indicating that F1 values change only minimally from steady-state to vowel offset. Given that the /i/ steady-state values for F1 are already very low (mean 346.18), this lack of change is not surprising. The mean F1 ratio for lax /ɪ/, on the other hand, is a fraction, indicating that F1 values fall from steady-state to onset, reflecting the fact that alveolar and labial codas do typically exert formant-lowering effects. The mean F2 steady-state of tense /i/ is greater than that of lax /ɪ/, in agreement with previous findings (e.g., Stevens & House, 1963). Finally, the mean F2 ratio values are fractions that are roughly

Table 1: Acoustic measurements (means, with standard deviations in parentheses) for the stimulus vowels used in Experiment 1. Transition ratio equals the duration of F2 formant transition divided by the total duration of vowel. F1 ratio equals the F1 at vowel offset divided by the F1 at steady-state. F2 ratio equals the F2 at vowel offset divided by the F2 at steady-state. For measurement details, see Pycha (2016).

	Transition ratio	Duration (ms)	F1 steady-state (Hz)	F1 ratio	F2 steady-state (Hz)	F2 ratio
Tense /i/	0.20 (0.13)	175 (51)	346 (52)	1.01 (0.28)	2773 (253)	0.80 (0.15)
Lax /ɪ/	0.77 (0.11)	146 (46)	570 (69)	0.79 (0.16)	2196 (179)	0.86 (0.12)

equivalent across tense /i/ (0.80) and lax /ɪ/ (0.86) conditions. This is in accordance with our stimulus selection procedures, in which we attempted to match F2 spectral change (as a proportion of steady-state F2) across tense and lax conditions.

All tokens were amplitude-normalized. The 192 target stimuli were divided into two lists of 96 tokens, each containing four tense and four lax vowel tokens from each of the twelve speakers. We then added 72 filler tokens to each list. The filler items were produced by nine of the speakers. Half of the fillers contained the same vowels as the target items, but with coda consonants that were not stops (e.g., [gif] and [gɪf]), and half contained completely different vowels with sonorant codas (e.g., [zæn], [zɑɹ]).

2.1.2 Participants and procedure

Thirty-eight participants, undergraduates who were recruited through subject pool databases at the University of California, Davis and the University of Wisconsin, Milwaukee, completed Experiment 1. All were native speakers of American English with no reported hearing impairment.

Each participant performed the task in a quiet laboratory environment, seated in front of an individual computer equipped with a monitor, keyboard, mouse, and high-quality headphones. Each participant was randomly assigned to one of the two lists. Participants were informed that the average height of American women is 5 feet 4 inches (5' 4'') (Centers for Disease Control and Prevention [CDC], 2010). On a given trial, participants were presented with a single stimulus token and asked to estimate the height of the speaker. Participants typed height estimates with a keyboard, using the American convention of feet followed by inches (e.g., 5'03''). Each participant completed 168 trials (96 target words, plus 72 filler items) presented in a randomized order.

2.2 Results

We hypothesized that transition ratio should exert a reliable effect on speaker height estimates: Specifically, we predicted that longer formant-lowering transition ratios should correlate with taller height estimates. To test this hypothesis, we used linear mixed effects modelling, which allowed us to control for multiple other variables that might have influenced listener responses. The linear mixed-effects model was run in R using the *lmer()* function in the lme4 package (Bates, Maechler, Bolker, & Walker, 2016). Estimates for degrees of freedom, *t*-statistics, and *p*-values were computed using Satterthwaite approximation with the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2015).

The model included seven fixed effect predictors. First, Transition Ratio for each vowel (a continuous predictor) was included to test for a gradient effect of temporal extent of formant-lowering coarticulation on speaker height estimates. Mean Word Log f0 (the mean log f0 calculated across each vowel; continuous predictor) was included because fundamental frequency affects judgments of speaker height (e.g., Barreda & Liu, 2018). Coda Consonant Place of Articulation (a categorical predictor with two levels: Alveolar and Labial) was included because our target stimuli were balanced for containing a labial or alveolar coda. F1 Ratio (a continuous predictor) was included because it differed across tense /i/ versus lax /ɪ/ stimuli (see **Table 1**).

Maximum F2 value of the vowel (a continuous predictor) was included as well because steady-state F2 has been shown to affect judgments of speaker height (Barreda, 2016, 2017). Under the assumption that coarticulation lowers F2, the peak F2 value should reflect the canonical vowel target. However, F2 Max will be collinear with each vowel's calculated Transition Ratio: As the temporal extent of coarticulation increases, the ability for articulators to achieve the frontest position will be affected. Indeed, a simple linear regression on Transition Ratio and F2 Max

confirms high collinearity between these variables in the stimuli ($r = –.7, p < .0001$). Collinearity, or non-independence of predictors, is a problem for regression modeling because it violates the assumption of orthogonality of predictors. The result is that it is impossible to be sure that the model fitting procedure is correctly attributing variance to either predictor. However, collinearity between variables can be handled through residualization (Gorman, 2010; Zellou & Tamminga, 2014). Residualization is a method for orthogonalizing predictor variables (Kuperman, Bertram, & Baayen, 2008; Wurm & Fisicaro, 2014). One member of a pair of collinear predictors is taken as a baseline (here, Transition Ratio) and the other predictor (here, F2 Max) is regressed linearly on the values of the baseline. The values of the second predictor are then replaced in the model by the residuals of this regression, which are by definition strictly orthogonal to the baseline predictor values.

F2 Ratio (a continuous predictor) was also included in the model, because increased *magnitude* of coarticulation (rather than temporal extent, which is the focus of our hypothesis) is another factor that might also influence listeners' failure to compensate for coarticulation (see, for example, Zellou, 2017). F2 Ratio is also collinear with Transition Ratio in the stimuli ($r = –.4, p < .001$). Therefore, we residualized F2 Ratio by Transition Ratio. Vowel Duration (a continuous predictor) was included because the stimuli were collected from various prosodic positions, such that vowel duration varied from one token to the next. While there is no existing literature to suggest that there is an effect of vowel duration on speaker height judgments, listeners' perceptions are generally influenced by vowel duration (e.g., Pisoni, 1973). Vowel Duration was also collinear with Transition Ratio ($r = –.2, p <. 001$), a negative correlation suggesting that longer transition ratios tended to occur on shorter vowels. Therefore, we residualized Vowel Duration by Transition Ratio.

Height responses were converted to centimeters. The model included by-participant and by-speaker random intercepts. All continuous variables, including the dependent variable, were centered and scaled (prior to residualization). The model output is provided in **Table 2**.

The key result is that Transition Ratio significantly predicted speaker height estimates. As the positive coefficient (0.04) indicates, as transition ratio increases (i.e., greater temporal extent of F2-lowering coarticulation), so do estimates of speaker height. Two additional predictors also exerted an effect. As expected, f0 had a reliable effect on height judgments: Higher f0 led to smaller height estimates. There was a main effect of Vowel Duration, as well, whereby longer vowels were judged to be produced by taller speakers. No other main effects were significant.

Table 2: Summary statistics from the lmer model run on speaker height judgments (Experiment 1).

	Est.	*SE*	df	*t*-value	*p*-value	
(Intercept)	0.01	0.12	29	0.06	0.95	
F0	−0.07	0.03	1466	−2.66	0.01	**
Vowel Duration	0.07	0.02	2933	3.51	<.001	***
Transition Ratio	0.04	0.02	3489	2.60	0.01	**
F1 Ratio	−0.03	0.02	3485	−1.69	0.09	
Coda POA [Labial]	−0.02	0.04	3486	−0.43	0.67	
F2 Ratio	0.00	0.02	3443	−0.08	0.94	
Max F2	−0.06	0.03	1954	−1.89	0.06	.

The effect of Transition Ratio on speaker height judgments for each vowel is illustrated in **Figure 2**. As seen, the effect of Transition ratio (x-axis) on speaker height estimates (y-axis) is very robust for the lax vowel /ɪ/ (left panel). Yet, we see a minimal influence of Transition ratio on reported speaker height for tense /i/ (right panel).

To test whether there were indeed vowel-specific effects of coarticulation on speaker height estimates, two post-hoc linear mixed effect regression models were run separately on lax /ɪ/ and tense /i/. These subset models contained the same fixed-effects and random-effects structure as the original model. Results are in **Table 3**. Confirming what the data suggest in **Figure 2**, the effect of Transition Ratio was significant within lax /ɪ/, but non-significant within tense /i/. Vowel Duration and F2 Max also predict speaker height judgments for lax /ɪ/, but not for tense /i/, where only f0 exerted a significant effect.

2.3 Interim summary

The results of Experiment 1 reveal, broadly, that the interval between source and acoustic effect of variation can modulate how listeners build a model of the speaker. Specifically, in gɪC tokens with lax /ɪ/, there was a gradient effect whereby tokens with more temporally extensive coarticulation received taller estimates of speaker height. This effect was above and beyond any effect (or lack thereof) of overall lowering of F2 (i.e., magnitude of coarticulation) or other spectral properties.

This result is in line with previous findings showing that listeners interpret lowered formants as evidence of a taller speaker (Fitch, 1994; Ives et al., 2005; Rendall et al., 2007; Smith & Patterson, 2005; Smith, Patterson, Turner, Kawahara, & Irino, 2005) and also broaden these findings in two ways. First, we have shown that listeners make height attributions even when the lowered formants originate from a coda consonant, rather than from speakers themselves (Rendall et al., 2007) or from intrinsic vowel quality (Barreda, 2016, 2017). Second, we have shown that listeners make height attributions in a gradient fashion, correlating with increased temporal extent of coarticulation. This is a notable result because it suggests that failures to compensate are not arbitrary or idiosyncratic events, but rather follow a lawful pattern: When coarticulatory effects increase their temporal reach, the likelihood of misattributing acoustic influences from a source on a target sound increases.

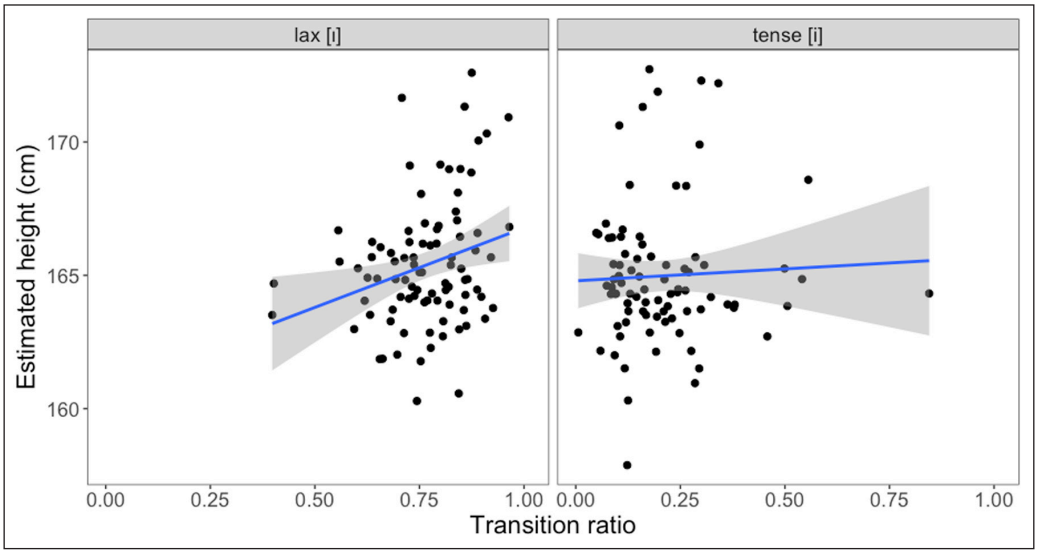

Figure 2: Experiment 1. Estimated mean speaker height judgment for each item, by transition ratio (duration of lowered F2 formant transition) for target stimuli with lax /ɪ/ and tense /i/ vowels. Ribbons represent the 95% confidence interval.

Table 3: Summary statistics from two lmer models run on speaker height judgments from lax /ɪ/ (top) and tense /i/ (bottom) (Experiment 1).

lax /ɪ/ subset model	Est.	*SE*	df	*t*-value	*p*-value	
(Intercept)	−0.19	0.14	107.5	−1.36	0.18	
F0	0.02	0.04	274	0.44	0.66	
Vowel duration	0.13	0.04	1107	3.21	<.001	***
Transition Ratio	0.25	0.11	495.7	2.35	0.02	**
F1 ratio	0.00	0.05	992.6	0.04	0.97	
Coda POA [Labial]	−0.03	0.08	1390	−0.42	0.67	
F2 Ratio	0.04	0.05	1022	0.81	0.42	
Max F2	−0.24	0.06	147.8	−3.69	<.001	***
tense /i/ subset model	**Est.**	***SE***	**df**	***t*-value**	***p*-value**	
(Intercept)	0.07	0.16	36.10	0.46	0.65	
F0	−0.16	0.04	630.60	−4.04	<.001	***
Vowel duration	0.01	0.03	1684.00	0.20	0.84	
Transition Ratio	0.09	0.07	987.40	1.23	0.22	
F1 ratio	−0.04	0.02	1714.00	−1.86	0.06	.
Coda POA [Labial]	−0.04	0.05	1692.00	−0.80	0.42	
F2 Ratio	0.03	0.03	1662.00	1.01	0.31	
Max F2	−0.03	0.06	376.70	−0.50	0.62	

However, the results of Experiment 1 also raise some issues that require further investigation. The most pressing question is, why do we see a gradient correlation between transition ratio and height judgments for lax /ɪ/, but not for tense /i/? Note that it is not the case that listeners displayed less variability in their height responses to /i/ than for /ɪ/. Rather, listeners' speaker height judgments for /ɪ/, but not for /i/, can be explained by temporal extent of coarticulation. There are two possible explanations that we will consider. First, it is possible that, in order for listeners to make a height misattribution, they need to hear a vowel whose transition ratio is at or above some minimum threshold value. Tokens with /i/ may not have reached that minimum value, particularly since we selected our stimuli so as to maximize the difference in transition ratios between tense /i/, where the values were comparatively small, and lax /ɪ/, where the values were comparatively large. A second possibility is that there is something about tense /i/ which makes it less likely that listeners will attribute coarticulatory-induced formant-lowering to speaker height. Vowel tenseness could be at play: Several studies report that lax vowels are produced with more extensive coarticulatory influences from adjacent consonants than corresponding tense vowels (Hoole & Mooshammer, 2002, for German; Pycha, 2016, for English).

The status of /i/ as a corner vowel could also be a contributing factor. Some studies have found that /i/, along with other corner vowels that overlap less with the other vowels, requires greater change in formant values to yield a different estimate of speaker height than interior vowels (Nearey, 1978). This scenario would mean that differences in the temporal extent of coarticulation for /i/ would not lead to different speaker size estimates, since the degree of spectral change is still insufficient to influence the perception of /i/. We aim to investigate these possibilities in Experiment 2a by extending the range of variation in the extent of coarticulation in both tense /i/ and lax /ɪ/, such that their

distributions have greater overlap. If the gradient effect for tense /i/ emerges, this would support a threshold prediction that the temporal extent of coarticulation needs to extend past a certain point into the steady-state vowel portion in order to affect the listeners' model of the speaker.

3. Experiment 2a: The effect of temporal extent of coarticulation on estimates of speaker height, using vowels without maximal differences

Experiment 2a was identical to Experiment 1, except that we selected stimuli so as to cover a wider range of transition ratio values, within both the tense /i/ and lax /ɪ/ categories.

3.1 Method

3.1.1 Stimuli

The 180 target stimuli for Experiment 2a were selected from the same corpus as those for Experiment 1, in which fifteen speakers produced the nonce words [git, gid, gip, gib] and [gɪt, gɪd, gɪp, gɪb]. Compared to Experiment 1, there were three differences in our selection procedure. First, we narrowed the set of target speakers to nine, eliminating three speakers for whom we did not have sufficient filler items containing the vowel /ɑ/. Second, our primary goal was now to select tokens whose transition ratios covered a wide range of values within both the tense /i/ and lax /ɪ/ conditions. Third, because the results of Experiment 1 did not reveal that coda place-of-articulation affects listeners' height judgments, we did not balance our stimuli for this factor. Crucially, we still balanced the tokens for speaker; thus, for each of these nine speakers, we examined the productions she had provided for gɪC tokens and selected ten that spanned the range from the smallest to largest transition ratio. We repeated the same process for each speaker for giC tokens. The acoustic characteristics of the vowels in the selected items are provided in **Table 4**.

Note that for transition ratio, our selection strategy for Experiment 2a primarily affected tokens with tense /i/. Whereas in Experiment 1 these tokens yielded a mean transition ratio of 0.20 and a standard deviation of 0.13 (see **Table 1**), in Experiment 2a they yielded a mean transition ratio of 0.33 and a standard deviation of 0.24—that is, an overall higher set of transition ratio values with greater variability. By comparison, our selection strategy had relatively little effect on tokens with lax /ɪ/. In Experiment 1 these tokens yielded a mean of 0.77 and a standard deviation of 0.11, while in Experiment 2a they yielded a mean of 0.76 and a standard deviation of 0.13. Presumably, this reflects a natural constraint on speakers' productions of lax /ɪ/ in CVC contexts. The remaining acoustic characteristics are largely similar to those for Experiment 1.

The 72 filler tokens, eight from each of the nine speakers, were the same as those used in Experiment 1.

Table 4: Acoustic measurements (means, with standard deviations in parentheses) for the stimulus vowels used in Experiments 2 and 3. Transition ratio equals the duration of F2 formant transition divided by the total duration of vowel. F1 ratio equals the F1 at vowel offset divided by the F1 at steady-state. F2 ratio equals the F2 at vowel offset divided by the F2 at steady-state. For measurement details, see Pycha (2016).

	Transition ratio	Duration (ms)	F1 steady-state (Hz)	F1 ratio	F2 steady-state (Hz)	F2 ratio
Tense /i/	0.33 (0.24)	152 (42)	352 (38)	0.97 (0.13)	2699 (223)	0.86 (0.09)
Lax /ɪ/	0.76 (0.13)	136 (42)	560 (69)	0.82 (0.13)	2204 (162)	0.86 (0.08)

3.1.2 Participants and procedure

Forty-two participants, University of California, Davis undergraduates, recruited through the Psychology Department subject pool, completed Experiment 2a. None of them had participated in Experiment 1. All were native speakers of English with no reported hearing impairment. Data from two participants was lost due to technical failure.

Three lists were created, each containing stimuli from six of the nine speakers. Assignment of speakers to a list was counterbalanced so that each speaker was assigned to two lists. Participants were randomly assigned to a list. Participants completed 240 experimental trials (4 vowel types × 10 tokens × 6 speakers), presented in randomized order. The procedure was the same as described for Experiment 1.

3.2 Results

Statistical analysis followed the same procedure as in Experiment 1, except that the model for Experiment 2a included the additional predictor of Vowel Identity (a categorical predictor with two values, tense /i/ and lax /ɪ/). As before, Transition Ratio, F2 Max, F2 Ratio, Mean Word Log F1, F1 Ratio, and Vowel Duration were all included, as described for Experiment 1. F2 Max and F2 Ratio were both collinear with Transition Ratio. Therefore, F2 Max was residualized by Transition Ratio, and F2 Ratio was also residualized by Transition Ratio. We also residualized Vowel Duration by Transition Ratio.

The model included four two-way interactions. Critically, Transition Ratio and Vowel were included to test whether temporal extent of coarticulation predicts estimated speaker height differently for lax /ɪ/ vs. tense /i/ vowels. The interaction of F2 Max and Vowel, as well as the interactions of F2 Ratio and Vowel, and F1 Ratio and Vowel, were included to test whether these acoustic values differently influenced within-vowel judgments of speaker height. All continuous variables, including the dependent variable, were centered and scaled (prior to residualization). By-participant and by-speaker random intercepts were included in the model. The model output is provided in **Table 5**.

The key result is a significant main effect of Transition Ratio, with a positive coefficient estimate indicating that the estimated speaker height increases with the temporal extent of coarticulation. In addition to this, the interaction between Transition Ratio and Vowel

Table 5: Summary statistics from the lmer model run on speaker height judgments (Experiment 2a).

	Est.	*SE*	df	*t*-value	*p*-value	
(Intercept)	−0.14	0.16	57	−0.91	0.37	
F0	0.01	0.01	4508	0.59	0.55	
Vowel [tense /i/]	0.05	0.11	4499	0.45	0.65	
Transition Ratio	0.07	0.03	4205	2.53	0.01	**
F1 Ratio	0.08	0.11	4500	0.72	0.47	
Vowel Duration	0.03	0.02	4418	1.85	0.06	.
F2 Ratio	−0.01	0.01	4505	−0.54	0.59	
Max F2	−0.03	0.03	1631	−1.00	0.32	
Vowel*Transition ratio	0.06	0.03	4511	2.28	0.02	**
Vowel*F1 ratio	−0.15	0.11	4513	−1.42	0.16	
Vowel*F2 ratio	0.01	0.01	4513	0.98	0.33	
Vowel*Max F2	−0.02	0.02	4496	−0.83	0.41	

was significant. As illustrated in the left panel of **Figure 3** for lax /ɪ/, listeners associate greater speaker heights with increasing temporal extent of coarticulation. As the right panel shows, this effect is not observed for tense /i/. In fact, although the estimated regression slope for tense /i/ appears negative, a post-hoc linear mixed effect regression model run just on tense /i/ (including the same fixed- and random-effects structure as the original model, except Vowel was not included) reveals a non-significant effect for Transition Ratio ($t(2205) = 1$, $p = .3$). An identical model run just on lax /ɪ/ confirms a significantly positive effect of Transition Ratio on speaker height estimates ($t(2214) = 3.2$, $p < .001$). No other effect was significant at the $p < .05$ level in either post-hoc model.

3.3 Interim summary

Our findings for Experiment 2a are similar to those of Experiment 1. Again, we see that the temporal distance between source and effect of variation can modulate how listeners build a model of the speaker. Specifically, in gɪC tokens with lax /ɪ/ there is a gradient effect whereby tokens with more temporally extensive coarticulation received taller estimates of speaker height. This effect was above and beyond any effect (or lack thereof) of overall lowering of F2 (i.e., degree of coarticulation) or other spectral properties.

A puzzling aspect of Experiment 1 results concerned the difference in behavior between vowel-phonemes. Specifically, why was there no correlation between transition ratio and height judgments for tense /i/? We had speculated that this null result may have originated from our stimulus selection procedure, but the results of Experiment 2a, in which we followed a different procedure, show that this was not the case. We revisit this issue in the General Discussion (Section 5).

4. Experiment 2b: The effect of temporal extent of coarticulation on vowel-phoneme identification

Both Experiment 1 and Experiment 2a demonstrate that increasing temporal extent of formant-lowering coarticulation on lax /ɪ/ leads to listener judgments of a taller speaker. This result suggests that the temporal distance between the coarticulatory source of variation and its effect modulates the models that listeners build of speakers. Does temporal distance also affect vowel-phoneme perception? In American English, lax /ɪ/ and the vowel /ʌ/ (as in

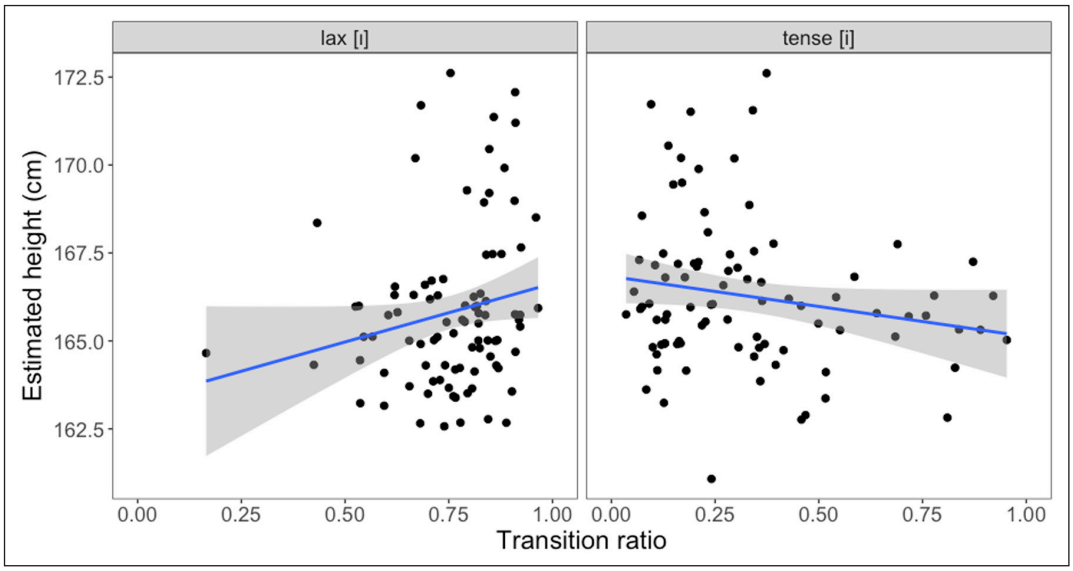

Figure 3: Experiment 2a. Mean estimated speaker height judgments for each item, by transition ratio (duration of lowered F2 formant transition) for target stimuli with lax /ɪ/ and tense /i/ vowels. Ribbons represent the 95% confidence interval.

hut or *hub*) have similar vowel heights, and differ primarily in backness, with /ʌ/ having an intrinsically lower F2 value than /ɪ/. Therefore, failure to compensate for lowered F2 might lead listeners to interpret the vowel identity of lax /ɪ/ as /ʌ/. Specifically, if listeners fail to compensate for temporally extensive coarticulation on lax /ɪ/, we predict a greater likelihood of listeners interpreting the intended vowel quality as a central /ʌ/.

Experiment 2b tests this prediction using a vowel discrimination task. In critical trials, listeners heard natural productions of gɪC nonce words ([gɪp, gɪb, gɪt, gɪd]) as well as two synthesized, isolated vowels: one with steady-state formant structure typical for /ɪ/ and one with formant structure typical for /ʌ/. Listeners selected the synthetic sound that they thought sounded most similar to the vowel in the gɪC nonce word. Tokens were the same as those from Experiment 2a, with a wide range of values for the temporal extent of coarticulation. If listeners fail to perceptually compensate, they will be more likely to interpret /ɪ/ as /ʌ/. Since there was no effect for tense /i/ in Experiment 2a, we likewise predicted that there would not be an effect of temporal extent of coarticulation for this vowel in Experiment 2b.

4.1 Method

4.1.1 Stimuli

Stimuli were the same tokens used in Experiment 2a. We used 90 gɪC tokens with lax /ɪ/ (ten from each of nine female speakers), and 90 giC tokens with tense /i/ (again, ten from each of nine female speakers). Fillers included 10 CVC tokens with [ʌ] and 10 CVC tokens with [ɑ] from each of these speakers.

In addition to the natural gVC tokens, four synthesized vowel-phonemes, representing canonical /i, ɪ, ʌ, ɑ/, were created. A single, isolated /ɪ, i, ʌ, ɑ/ vowel was synthesized separately for each speaker using the mean vowel pitch and vowel duration from that speaker's natural tokens (provided in **Table 6**). While the synthetic vowels were created to have the mean pitch and duration properties for each individual speaker, the formant properties needed to represent canonical, steady-state American English vowels. Therefore, F1 and F2 values for each vowel-phoneme were based on average productions from adult female speakers of American English (Hillenbrand, Getty, Clark, & Wheeler, 1995) provided in **Table 7**. F3 was fixed across vowel categories at an appropriate value for an adult female speaker, following synthetic vowel conventions (Klatt, 1980). All vowels had steady-state formant frequencies and were synthesized using a Klatt-style synthesis program.

Table 6: Mean pitch and duration for the vowels from tokens produced by nine female speakers' productions used in Experiment 2b.

	S1	S2	S4	S5	S6	S7	S8	S10	S11
F0 (Hz)	170	180	210	220	190	190	200	230	190
Duration (ms)	140	160	150	120	120	150	170	150	120

Table 7: First three formant frequencies for synthetic vowels created for Experiment 2b.

	/i/	/ɪ/	/ʌ/	/ɑ/
F1	440	400	700	950
F2	2800	2500	1500	1400
F3	3000	3000	3000	3000

4.1.2 Participants and procedure

The same 42 listeners who participated in Experiment 2a completed an AXB discrimination task. On each trial, listeners heard two synthetic vowels (i.e., "A" and "B") flanking a naturally produced gVC token (i.e., "X"). Listeners were instructed to indicate whether the first or the last sound matched the vowel in the nonsense word that they heard in each trial.

Synthetic-vowel pairings were the same within vowel categories, but different across vowel categories. Critically, lax /ɪ/ stimuli were flanked with synthetic /ɪ/ and synthetic /ʌ/ vowels to test whether greater temporal extent of coarticulation increases the likelihood of misperceiving /ɪ/ as /ʌ/. If listeners fully compensate for the effects of coarticulatorily lowered formants, matching a gɪC token with /ʌ/ should be unlikely. However, if listeners do not compensate, matching a gɪC token with /ʌ/ should become more likely.

In American English, the F2 differences between /i/ and /ɪ/ are roughly comparable to those between /ɪ/ and /ʌ/. That is, tense /i/ and lax /ɪ/ have similar vowel heights, and differ primarily in backness, with /ɪ/ having an intrinsically lower F2 value than /i/. Therefore, failure to compensate for lowered F2 might lead listeners to interpret the vowel identity of tense /i/ as /ɪ/. Therefore, the giC stimuli were flanked by synthetic /i/ and /ɪ/ vowels, which would allow us to evaluate whether the same effect is present for tense /i/. Filler trials consisted of CʌC items flanked by /ʌ/ and /ɑ/, as well as CɑC items flanked by /ɑ/ and /i/, so that an equal proportion of each synthetic vowel sound occurred throughout testing.

Participants were presented with the same set of stimuli (from the same set of six speakers) assigned to them in Experiment 2a. Participants completed 240 discrimination trials (4 vowel types × 10 tokens × 6 speakers), in random order. Ordering of flanking vowels (either, e.g., AXB or BXA) was assigned equally within vowel types in a list and counterbalanced across lists. Within-trial inter-stimulus intervals (ISIs) between flanking vowels and the token were 500 milliseconds. The inter-trial interval (ITIs) was 1 second.

4.2 Results

Overall vowel discrimination performance was higher for gɪC tokens (74% mean percent /ɪ/ response) than for giC tokens, which was at chance (50% /i/ response). While we had no predictions about differences across vowels with respect to overall performance, the decrease in accuracy for /i/ tokens possibly reflects the acoustic similarity between the synthesized vowels (as seen in **Table 7**, synthesized /i/ and /ɪ/ were only a few hundred Hz apart in F2); also, the inherent duration differences between these vowels was not reflected in the synthetic stimuli and possibly contributed to the ambiguity between these sounds.

Figure 4 provides listeners' averaged responses to AXB discrimination trials, by transition ratio (x-axis), separately for gɪC (left panel) and giC (right panel). As seen, the accuracy in vowel discrimination for gɪC decreases as transition ratio increases. This aligns with our prediction: As temporal extent of coarticulatory-induced F2-lowering on lax /ɪ/ increases, the likelihood of interpreting the vowel as /ʌ/ increases. Meanwhile, mirroring what was observed with these stimuli in Experiment 2a, the temporal extent of coarticulation on tense /i/ appears to not predict differences in vowel identification responses.

To test the statistical significance of the patterns observed in **Figure 4**, discrimination responses for tense /i/ and lax /ɪ/ (1 = intended vowel-phoneme identification, 0 = other option) were fitted to a mixed effects logistic regression model. The model was run in R using the *glmer()* function in the lme4 package (Bates et al., 2016). The fixed-effects structure of the model was identical to that described for Experiment 2a (except that f0 was not included). There were five fixed effect predictors: Transition Ratio, Vowel,

Vowel Duration, F2 Max, and F2 Ratio. The interaction of Transition Ratio and Vowel was included to test whether temporal extent of coarticulation predicts discrimination performance differently for lax versus tense vowels. By-participant and by-speaker random intercepts were included in the model, as well. The model output is shown in **Table 8**.

A post-hoc logistic mixed effect regression model run just on tense /i/ (including the same fixed- and random-effects structure as the original model, except Vowel was not included) showed that vowel identification was not significantly affected by Transition Ratio ($z = -.6$, $p = 0.5$). An identical model run just on lax /ɪ/ confirms a significantly positive effect of Transition Ratio on vowel identification responses ($z = -2.2$, $p < .05$). In addition, both models computed a significant effect of F2 Max: As steady-state F2 was higher, listeners were more likely to correctly identify the intended vowel ($z = 2.2$, $p < .05$ for tense /i/; $z = 3.3$, $p < .001$ for lax /ɪ/). No other effect was significant at the $p < .05$ level in either post-hoc model.

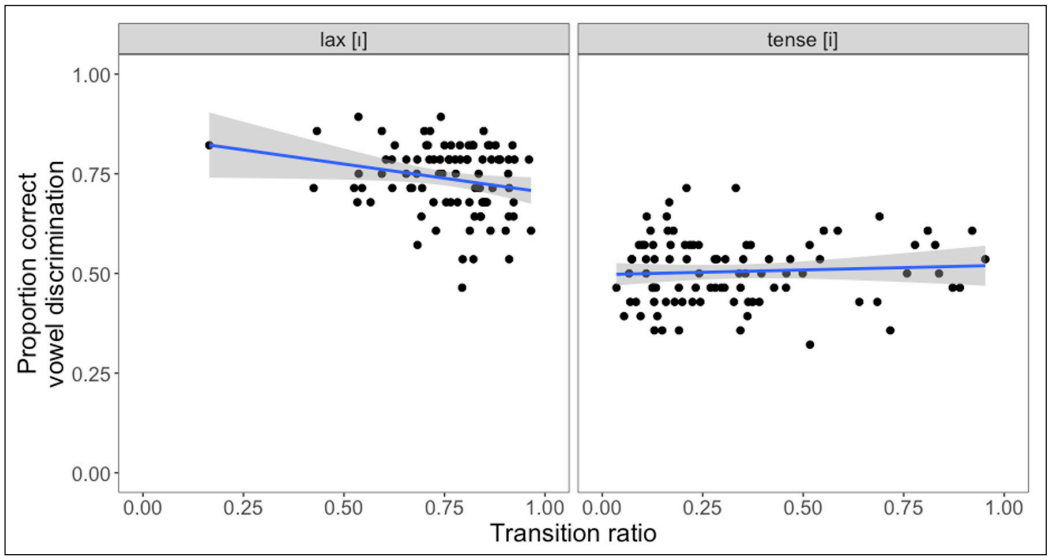

Figure 4: Experiment 2b. Mean vowel phoneme discrimination performance for each item, by transition ratio (duration of lowered F2 formant transition) for target stimuli with lax /ɪ/ (left panel) and tense /i/ (right panel) vowels. Ribbons represent the 95% confidence interval.

Table 8: Summary statistics from the glmer model run on vowel-phoneme discrimination (Experiment 2b).

	Est.	*SE*	*z*-value	*p*-value	
(Intercept)	0.68	0.11	6.14	<.001	
Vowel [tense /i/]	0.72	0.07	10.31	<.001	***
Transition Ratio	−0.16	0.07	−2.48	0.01	**
F1 Ratio	0.00	0.04	0.08	0.94	
Vowel Duration	−0.04	0.04	−1.10	0.27	
F2 Ratio	0.05	0.03	1.31	0.19	
Max F2	0.17	0.05	3.60	<.001	***
Vowel*Transition ratio	−0.12	0.07	−1.83	0.05	.
Vowel*F1 ratio	0.02	0.04	0.42	0.67	
Vowel*F2 Ratio	−0.01	0.03	−0.23	0.82	
Vowel*Max F2	0.05	0.05	1.11	0.27	

4.3 Discussion

Our findings for Experiment 2b demonstrate that the size of the temporal gap between source and acoustic effect can modulate how listeners perceive vowel-phonemes. Specifically, in gɪC tokens with lax /ɪ/, there is a gradient effect whereby tokens with more temporally extensive coarticulation were more likely to be perceived as /ʌ/.

5. General discussion

The over-arching goal of this study was to further explore the circumstances under which listeners attribute the acoustic effect of coarticulation to a different source. To that end, we investigated how variation in the temporal extent of coarticulatory-induced formant lowering in vowels influenced two different listener behaviors: speaker-model representation and vowel-phoneme perception. Experiment 1 asked listeners to estimate speaker height based on giC and gɪC tokens produced by 12 female speakers. Results indicated a gradient effect: Within lax /ɪ/, greater temporal extent of coarticulation correlated with greater height judgments. Experiment 2a was similar, except that temporal extent of coarticulation varied across a wider range of values than in Experiment 1. Results again indicated a gradient effect: Within lax /ɪ/, greater temporal extent of coarticulation correlated with greater height judgments. Taken together, Experiments 1 and 2a demonstrate that the temporal extent of coarticulation affects listeners' speaker-model representations, by modulating their height estimates. A related question is whether the same factors influence vowel-phoneme perception. To that end, Experiment 2b used the same tokens and the same participants from Experiment 2a, but this time we asked listeners to perform an AXB discrimination task. Results showed that, within lax /ɪ/, greater temporal extent of coarticulation correlated with greater likelihood of judging the vowel to be /ʌ/. Thus, Experiment 2b demonstrates that the temporal extent of correlation also affects listeners' interpretation of vowel identity. We discuss the implications of these results for listener-based theories of sound change.

5.1 Implications for sound change: The distance between source and acoustic effect

A key motivation for the current study was Ohala's speculation that acoustic effects of coarticulation which are temporally more distant from their source should be more susceptible to misattribution. Specifically, he speculated that "…the farther away the conditioning environment is from the conditioned change—that is, the greater the temporal gap between cause and effect—the more difficult it will be for the listener to be able to establish the causal link between the two and use this link as the basis for correction" (1993, p. 247). For at least the lax vowel /ɪ/, the results of the current study clearly support this speculation: The greater the temporal extent of coarticulation, the more listeners were likely to fail to attribute that variation to the intended source. This result held for both types of source attribution that we investigated, vowel-phoneme perception and speaker-models.

Our results also support and extend previous findings in the compensation literature. We noted that Beddor and colleagues (Beddor et al., 2013) had previously investigated the temporal relationship between source and acoustic effect. Their stimulus words, such as *bend* and *bent*, had two categorically-opposing values for the temporal extent of nasal coarticulation on the vowel, early versus late, and their results showed that listeners did respond to these two values of coarticulation in different ways. Our results extend this finding in three ways. First, we show that the temporal relationship between source and acoustic effect is relevant not just for nasality, but also for a different type of coarticulation, namely formant-lowering transitions triggered by a coda consonant. Second, the temporal extent of coarticulation in our stimuli varied gradiently over a range of values, rather than

between two opposing categories, and our results showed that listeners were sensitive to this gradience. Third, we showed that the interval between coarticulatory source and its effect is important not just for vowel-phoneme perception, but also for speaker-models.

Our results also support and extend previous findings in the speaker-models literature. We noted that many studies had previously demonstrated a relationship between vowel formant values and judgments of speaker height (Fitch, 1994; Ives et al., 2005; Rendall et al., 2007; Smith & Patterson, 2005; Smith et al., 2005). These studies all manipulated steady-state formant values. To our knowledge, the current study is the first one to manipulate formant transitions, rather than steady-states. It is also the first to manipulate the (relative) duration, rather than spectral magnitude, of formant values. The implication is that speaker-models can be affected by a broader set of factors than previously thought: Not only are across-the-board changes in formant values important, but so are coarticulatory perturbations created by contextual variation, which are ubiquitous in speech production.

5.2 Implications for sound change: Phoneme-specific patterns

Across three experiments, our results indicate that listeners respond differently to variations in temporal extent of coarticulation present in tense /i/ versus lax /ɪ/. In contrast to lax /ɪ/, listeners' responses to tense /i/ cannot be explained by any measure of coarticulation we included in our analyses. In other words, while source-misattribution can explain the systematic patterns of responses for lax /ɪ/, it does not explain response patterns to tense /i/ stimuli. While this issue cannot be fully resolved here, we consider two possible explanations for this vowel-specific effect.

First, the vowel-specific effect may derive from fundamental differences in coarticulatory patterns across vowels. For example, the production data of Pycha (2016) showed that for tense /i/, the coarticulatory effects of a following coda consonant were confined to a region immediately adjacent to that consonant, but for lax /ɪ/, these effects extended much further into the vowel. In other words, despite the fact that lax /ɪ/ is phonemic in American English, it nevertheless appears to be underspecified relative to tense /i/—that is, its formant trajectories (particularly for F2) are characterized not by a steady-state period, but primarily by interpolation toward the values exerted by the adjacent consonant(s). Conceivably then, the important difference between /i/ and /ɪ/ is that they differ in the relative time at which coarticulation due to the following consonant exerts its influence on the vowel. If this is the case, our results suggest that gradient source-misattribution occurs only for vowels like /ɪ/ for which the temporal extent of coarticulation is large. Future studies could test this prediction by identifying other underspecified vowels and investigating whether they behave similarly to /ɪ/ (or, conversely, more perceptual resistance to coarticulation in vowels that behave similarly to /i/).

Alternatively, the vowel-specific patterns might derive from the location of tense /i/ in the vowel space. A sliding template model of vowel normalization proposes that vowel categorization involves matching the input to the closest phoneme after shifting a vowel space template along an F1 × F2 dimension based on the listener's estimate of speaker size (Nearey & Assmann, 2007). Such a model predicts, based on the shape of the vowel space, that the interior vs. peripheral vowel categories would be differentially susceptible to mistakes in phoneme mapping. Interior vowels (such as /ɪ/ and /ʌ/) can overlap in a shifting vowel space. But, because of the position of tense /i/, essentially in a peripheral corner of the vowel space, it is less likely to overlap with another vowel category in a shifting vowel space. Indeed, previous studies have shown that tense /i/ is more resistant to misperception than other, interior, vowels. Nearey (1978) found that a 45% change in the formant frequencies of a synthetic /i/ was required to shift apparent size judgment

from an adult male to a child, while interior vowels required only a 20% shift. Evidence such as this suggests that the peripheral vowels /i a u/ might serve to calibrate a given speaker's vowel system and be used by listeners to guide the mapping of the other vowels (Ainsworth, 1975; Nearey, 1989). However, responses to speaker height and phoneme discrimination were just as variable for /i/ as for /ɪ/. While it is possible that the lack of a gradient effect for tense /i/ is due to its properties as a peripheral vowel and its general resistance to the influence of coarticulation on perceptual responses, further work is needed to understand what is conditioning the variability in /i/ seen across our data.

Future studies can shed light on these potential explanations. For example, if tenseness is the cause of such differences in coarticulatory timing, then our findings should extend to other pairs of tense-lax vowels, such as /u/ versus /ʊ/ and /e/ versus /ɛ/. On the other hand, if there is something particular about corner vowels, comparing other peripheral vowels to interior vowels should yield similar patterns to our /i/ versus /ɪ/ observations. Investigations into vowel-specific differences in the temporal extent of coarticulation and its perceptual ramifications is a promising avenue for future research.

5.3 Relationship between perceptual compensation failures for vowel identification and speaker-models

We began this study with the idea that perceptual compensation for vowel-phoneme perception and speaker-models both impose similar requirements on the listener, and our findings have demonstrated that both processes are modulated by the temporal extent of coarticulation. However, the precise relationship between vowel-phoneme perception and speaker-model construction remains an open question that cannot be directly addressed by our data. Further work is needed to understand their relationship. In the meantime, however, we can note several clear differences between these two processes, which have implications for models of sound change.

On the one hand, successful perceptual compensation would seem to depend primarily on accurate identification of the source of the coarticulatory perturbation, which typically occurs in a short time. For example, to compensate for the lowered formants in *heap* /hip/ or *hip* /hɪp/, it is important that the listener can actually perceive the /p/. In any situation where clear perception of /p/ is compromised, we would expect compensation to be either partial or non-existent (Kawasaki, 1986). Such situations could be fleeting, such as when a sound like /p/ is interrupted by background noises or coughing. Or they could be more systematic, such as when a segment like /p/ is (for whatever reason) inherently more difficult for listeners to perceive than other segments. Note that for any individual instance of misattribution, top-down information could serve as a corrective factor: If the listener can reconstruct the signal [hɪp] from top-down information provided by the lexeme /hɪp/, then this could conceivably restore the percept of the source /p/ (Samuel, 1996), even if the source is not physically present (Ohala & Feder, 1994).

On the other hand, successful speaker-models would seem to depend primarily on consistent and reliable identification of the characteristics of *a particular speaker*, namely the person producing the utterance, over a longer time window of their entire productions. For example, in order to adjust for vowel formants that may be somewhat lowered, it is important that the listener build a model of a speaker who is somewhat tall. In any situation where information about the speaker is compromised, we would expect inaccurate models (Eklund & Traunmüller, 1997; Johnson, 2008). One obvious example is visual information. If the listener cannot see the speaker, he or she is missing one cue to height. Similarly, if the speaker's characteristics do not conform to listener expectations—e.g., if the speaker has a voice that is non-stereotypical for their gender—the listener is missing another cue

to the speaker's identity. Both situations can lead to changes in perception of the speech signal (Johnson et al., 1999). Note that for any individual instance of misattribution, previous or subsequent information in that same speaker's utterance could serve as a corrective factor: If the global, non-local patterns of that speaker's vowel formants, e.g., either higher formant characteristics or information taken over a longer time window than one vowel, provide good information about his or her identity, this could conceivably adjust the percept of an individual vowel (Ladefoged & Broadbent, 1957).

Thus, the factors that influence these two types of source attribution do differ. In general, phoneme perception involves short-term stretches of the speech signal, while speaker-models can presumably be derived from longer aspects of the speaker's utterances. If this characterization is correct, it has implications for theories of sound change. We would predict that sound change arising from perceptual compensation should be linked to specific combinations of sounds, such as V + N or V + p; this connection has been previously explored in the literature (Beddor, 2010; Beddor et al., 2002, 2013; Blevins, 2004; Garrett & Johnson, 2013; Ohala, 1993). Meanwhile, sound change resulting from misattributing acoustic properties to apparent speaker characteristics are not attested as far as we know, though the issue remains relatively unexplored. Exploring the relationship, as well as the differences, between perceptual compensation during phoneme perception and speaker-model construction should be a fruitful avenue of future work.

5.4 Limitations

The interpretations of our results are constrained by certain limitations. We tested a single pair of vowels, tense /i/ and lax /ɪ/, from one language variety, American English. An important goal for future work will be to extend our hypotheses to other vowels, and also to consonants, beyond English. We tested these vowels specifically within the context of coda consonants, and not e.g., in simple CV syllables, so we do not know how listeners would judge these vowels in the absence of formant-lowering coarticulatory sources. This is a natural limitation for the lax vowel /ɪ/, which cannot occur in isolated CV syllables (e.g., [gɪ] is ill-formed), although future research might explore potential work-arounds to this problem. Also, because our tokens were excised from different contexts (medial versus final position in a phrase, focused versus non-focused), prosodic cues may have influenced listener judgments. We limited these effects by amplitude-normalizing our tokens and including mean pitch for each token as a random factor in our statistical analyses; but differences in overall pitch contour, as well as vowel duration, remained.

We presented listeners with nonsense words, and not real words. If, as some researchers have argued, lexical effects can modulate compensation for coarticulation (e.g., Elman & McClelland, 1988), then our decision has the advantage of eliminating this confounding factor and presenting listeners with an optimally simple compensation task. Yet the potential disadvantage is that our findings may not directly apply to everyday situations, where listeners encounter real words. We also presented listeners with productions from exclusively female speakers. One advantage of this decision is that, over the course of an experiment, participants heard a smaller range of F1 and F2 values than they would have if we had included male and/or child speakers, and their judgments for height and vowel identity may have been more sensitive as a result. An obvious disadvantage, though, is that we do not know if our findings extend across these other speaker types. Finally, we did not measure the actual height of our speakers, so we do not know the extent to which our participants' judgments correlate with the facts. However, it was not ultimately the purpose of this study to determine whether listeners' height judgments are accurate; instead, we showed that those judgments are influenced by factors that have nothing to do with actual height.

6. Conclusion

A single acoustic output can be the result of multiple different articulatory and physical parameters. Ohala (1993) proposed that such ambiguity in the speech signal can result in mismatches between the speaker's intent and listener's interpretation of the utterance. The goal of our research was to pinpoint when and why such mismatches occur. We have shown that mismatches are more likely for nonce words containing lax vowel /ɪ/, compared to those containing tense vowel /i/. This suggests that certain vowels are less resistant to the influence of coarticulation in misperception than others. We have also shown that temporal dynamics play a strong role in modulating perceptual compensation. For nonce words with lax vowel /ɪ/, the greater the temporal extent of coarticulation from the coda, the greater the likelihood and amount of misattribution to a different source. Following Ohala (1993), this suggests a concrete diagnostic: We predict that misattributions are more likely for any situation in which the effects of coarticulation are temporally more removed from the source that gives rise to them, compared to those in which its effects are strictly local. Pinpointing how and why hypocorrections occur should remain a major objective for the study of speech perception and sound change.

Acknowledgements

For helpful comments and feedback, we thank participants of the 4th Workshop on Sound Change at the University of Edinburgh, Santiago Barreda, Jonathan Harrington, and two anonymous reviewers.

Competing Interests

The authors have no competing interests to declare.

References

Ainsworth, W. 1975. Intrinsic and extrinsic factors in vowel judgments. In: Fant, G., & Tatham, M. (eds.), *Auditory Analysis and Perception of Speech*, 103–113. New York: Academic Press.

Barreda, S. 2016. Investigating the use of formant frequencies in listener judgments of speaker size. *Journal of Phonetics, 55*, 1–18. DOI: https://doi.org/10.1016/j.wocn.2015.11.004

Barreda, S. 2017. Listeners respond to phoneme-specific spectral information when assessing speaker size from speech. *Journal of Phonetics, 63*, 1–18. DOI: https://doi.org/10.1016/j.wocn.2017.03.002

Barreda, S., & Liu, Z. Y. 2018. Apparent-talker height is influenced by Mandarin lexical tone. *The Journal of the Acoustical Society of America*, *143*(2), EL61–EL66. DOI: https://doi.org/10.1121/1.5022156

Bates, D., Maechler, M., Bolker, B., & Walker, S. 2016. lme4: Linear mixed-effects models using Eigen and S4. (Version 1.1–7).

Beddor, P. S. 2010. A coarticulatory path to sound change. *Language, 85*(4), 785–821. DOI: https://doi.org/10.1353/lan.0.0165

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*(4), 591–627. DOI: https://doi.org/10.1006/jpho.2002.0177

Beddor, P. S., & Krakow, R. A. 1999. Perception of coarticulatory nasalization by speakers of English and Thai: Evidence for partial compensation. *The Journal of the Acoustical Society of America, 106*(5), 2868–2887. DOI: https://doi.org/10.1121/1.428111

Beddor, P. S., McGowan, K. B., Boland, J. E., Coetzee, A. W., & Brasher, A. 2013. The time course of perception of coarticulation. *The Journal of the Acoustical Society of America, 133*(4), 2350–2366. DOI: https://doi.org/10.1121/1.4794366

Blevins, J. 2004. *Evolutionary phonology: The emergence of sound patterns*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511486357

Byrd, D. 1992. Perception of Assimilation in Consonants Clusters: A Gestural Model. *Phonetica, 49*(1), 1–24. DOI: https://doi.org/10.1159/000261900

Eklund, I., & Traunmüller, H. 1997. Comparative study of male and female whispered and phonated versions of the long vowels of Swedish. *Phonetica, 54*(1), 1–21. DOI: https://doi.org/10.1159/000262207

Elman, J. L., & McClelland, J. L. 1988. Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language, 27*(2), 143–165. DOI: https://doi.org/10.1016/0749-596X(88)90071-X

Fitch, W. T. 1994. *Vocal tract length perception and the evolution of language*. Brown University. Retrieved from: https://pdfs.semanticscholar.org/2d70/0cc912117fadf974018bebcb9813a6b4d5ce.pdf.

Fitch, W. T., & Giedd, J. 1999. Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America, 106*(3), 1511–1522. DOI: https://doi.org/10.1121/1.427148

Garrett, A., & Johnson, K. 2013. Phonetic bias in sound change. In: Yu, A. (ed.), *Origins of sound change: Approaches to phonologization*, 51–97. Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199573745.003.0003

Gorman, K. 2010. The Consequences of Multicollinearity among Socioeconomic Predictors of Negative Concord in Philadelphia. *University of Pennsylvania Working Papers in Linguistics, 16*(2). Retrieved from: https://repository.upenn.edu/pwpl/vol16/iss2/9.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. 1995. Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America, 97*(5), 3099–3111. DOI: https://doi.org/10.1121/1.411872

Hoole, P., & Mooshammer, C. 2002. Articulatory analysis of the German vowel system. In: Auer, P., Gilles, P., & Spiekermann, H. (eds.), *Silbenschnitt und Tonakzente*, 129–152. Max Niemeyer Verlag. DOI: https://doi.org/10.1515/9783110916447.129

Ives, D. T., Smith, D. R. R., & Patterson, R. D. 2005. Discrimination of speaker size from syllable phrases. *The Journal of the Acoustical Society of America, 118*(6), 3816–3822. DOI: https://doi.org/10.1121/1.2118427

Johnson, K. 2008. Speaker normalization in speech perception. In: Pisoni, D., & Remez, R. (eds.), *The Handbook of Speech Perception*, 363–389. Blackwell.

Johnson, K., Strand, E. A., & D'Imperio, M. 1999. Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics, 27*(4), 359–384. DOI: https://doi.org/10.1006/jpho.1999.0100

Kawasaki, H. 1986. Phonetic explanation for phonological universals: The case of distinctive vowel nasalization. *Experimental Phonology*, 81–103.

Klatt, D. H. 1980. Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America, 67*(3), 971–995. DOI: https://doi.org/10.1121/1.383940

Kuperman, V., Bertram, R., & Baayen, R. H. 2008. Morphological dynamics in compound processing. *Language and Cognitive Processes, 23*(7–8), 1089–1132. DOI: https://doi.org/10.1080/01690960802193688

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. 2015. *Package 'lmerTest.'* R. Retrieved from: http://cran.uib.no/web/packages/lmerTest/lmerTest.pdf.

Ladefoged, P., & Broadbent, D. E. 1957. Information Conveyed by Vowels. *The Journal of the Acoustical Society of America, 29*(1), 98–104. DOI: https://doi.org/10.1121/1.1908694

Lieberman, P., & Blumstein, S. E. 1988. *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139165952

Mann, V. A. 1980. Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics, 28*(5), 407–412. DOI: https://doi.org/10.3758/BF03204884

Nearey, T. M. 1978. *Phonetic feature systems for vowels.* Indiana University. Retrieved from: https://elibrary.ru/item.asp?id=7214634.

Nearey, T. M. 1989. Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America, 85*(5), 2088–2113. DOI: https://doi.org/10.1121/1.397861

Nearey, T. M., & Assman, P. F. 2007. Probablistic sliding template models for indirect vowel normalization. In: Sole, M.-J., Beddor, P. S., & Ohala, M. (eds.), *Experimental Approaches to Phonology*, 246–269. OUP Oxford.

Niedzielski, N. 1996. Acoustic analysis and language attitudes in Detroit. *University of Pennsylvania Working Papers in Linguistics, 3*(1), 7.

Ohala, J. J. 1993. The phonetics of sound change. In: *Historical linguistics: Problems and perspectives*, 237–278.

Ohala, J. J., & Feder, D. 1994. Listeners' Normalization of Vowel Quality Is Influenced by 'Restored' Consonantal Context. *Phonetica, 51*(1–3), 111–118. DOI: https://doi.org/10.1159/000261963

Peterson, G. E., & Lehiste, I. 1960. Duration of Syllable Nuclei in English. *The Journal of the Acoustical Society of America, 32*(6), 693–703. DOI: https://doi.org/10.1121/1.1908183

Pisoni, D. B. 1973. Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics, 13*(2), 253–260. DOI: https://doi.org/10.3758/BF03214136

Pycha, A. 2016. Co-articulatory cues for communication: An investigation of five environments. *Language and Speech, 59*(3), 364–386. DOI: https://doi.org/10.1177/0023830915603878

Rendall, D., Vokey, J. R., & Nemeth, C. 2007. Lifting the curtain on the Wizard of Oz: Biased voice-based impressions of speaker size. *Journal of Experimental Psychology: Human Perception and Performance, 33*(5), 1208–1219. DOI: https://doi.org/10.1037/0096-1523.33.5.1208

Samuel, A. G. 1996. Does lexical information influence the perceptual restoration of phonemes? *Journal of Experimental Psychology: General, 125*(1), 28–51. DOI: https://doi.org/10.1037/0096-3445.125.1.28

Smith, D. R. R., & Patterson, R. D. 2005. The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *The Journal of the Acoustical Society of America, 118*(5), 3177–3186. DOI: https://doi.org/10.1121/1.2047107

Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H., & Irino, T. 2005. The processing and perception of size information in speech sounds. *The Journal of the Acoustical Society of America, 117*(1), 305–318. DOI: https://doi.org/10.1121/1.1828637

Stevens, K. N., & House, A. S. 1963. Perturbation of vowel articulations by consonantal context: An acoustical study. *Journal of Speech & Hearing Research, 6*(2), 111–128. DOI: https://doi.org/10.1044/jshr.0602.111

Wurm, L. H., & Fisicaro, S. A. 2014. What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language, 72*, 37–48. DOI: https://doi.org/10.1016/j.jml.2013.12.003

Yu, A. C. L. 2010. Perceptual compensation is correlated with individuals' "autistic" traits: Implications for models of sound change. *PLoS One, 5*(8), e11950. DOI: https://doi.org/10.1371/journal.pone.0011950

Yu, A. C. L., & Lee, H. 2014. The stability of perceptual compensation for coarticulation within and across individuals: A cross-validation study. *The Journal of the Acoustical Society of America, 136*(1), 382–388. DOI: https://doi.org/10.1121/1.4883380

Zellou, G. 2017. Individual differences in the production of nasal coarticulation and perceptual compensation. *Journal of Phonetics, 61*, 13–29. DOI: https://doi.org/10.1016/j.wocn.2016.12.002

Zellou, G., & Tamminga, M. 2014. Nasal coarticulation changes over time in Philadelphia English. *Journal of Phonetics, 47*, 18–35. DOI: https://doi.org/10.1016/j.wocn.2014.09.002

How to cite this article: Zellou, G. and Pycha, A. 2018 The gradient influence of temporal extent of coarticulation on vowel and speaker perception. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 9(1): 12, pp. 1–24, DOI: https://doi.org/10.5334/labphon.118

Submitted: 25 September 2017 **Accepted:** 20 April 2018 **Published:** 06 August 2018

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

OPEN ACCESS