

JOURNAL ARTICLE

Intonational variation and incrementality in listener judgments of ethnicity

Nicole Holliday¹ and Dan Villarreal^{2,3}

¹ Department of Linguistics and Cognitive Science, Pomona College, Claremont, CA, US

² New Zealand Institute of Language, Brain and Behaviour, University of Canterbury, Christchurch, NZ

³ Department of Linguistics, University of Pittsburgh, US

Corresponding author: Nicole Holliday (nicole.holliday@pomona.edu)

The current study examines how listeners make gradient and variable ethnolinguistic judgments in an experimental context where the speaker's identity is well-known. It features an open-guise experiment (Soukup, 2013) that assessed whether sociolinguistic judgments are subject to *incrementality*, with judgments increasing in magnitude as variable stimuli demonstrate more extreme differences. In particular, this task tested whether judgments of President Barack Obama as sounding 'more' or 'less' black (e.g., Alim & Smitherman, 2012) are sensitive to differences in intonation. Half of critical stimuli featured an L+H* pitch accent, which occurs more frequently in African American Language than in Mainstream U.S. English (Holliday, 2016). Four stimuli apiece were created from these phrases by making each pitch accent more extreme by semitone-based F0 steps. Seventy-nine listeners rated these stimuli via the question, "How black does Obama sound here?" Mixed-effects modeling indicated that listeners rated more phonetically extreme L+H* stimuli as sounding blacker, regardless of listener identity. A post-hoc analysis found that listeners attended to different voice quality features in L+H* stimuli. We discuss implications for research in intonation, ethnic identification, incrementality, language attitudes, and sociolinguistic awareness.

Keywords: Intonation; perception; sociophonetics; African American Language; language attitudes; variation

1. Introduction

Recent research in perceptual sociolinguistics has investigated a host of phonetic and phonological variables—primarily segmental—to assess the extent to which social meanings are constructed in perception, similar to the way they are constructed in ongoing production. Despite production research in sociolinguistics demonstrating how speakers use intonational variation to index various ethnic identities and social stances (Burdin, 2015; Holliday, 2016; Reed, 2016), there has been a general lack of perceptual research on the social meanings of intonational variables. In addition, while decades of research have demonstrated U.S. listeners' ability to distinguish African American and white voices (cf. Thomas & Reaser, 2004), these studies have also revealed challenges inherent in isolating speaker-specific variables that drive ethnic identification (Holliday & Jagers, 2015; Purnell, Idsardi, & Baugh, 1999); indeed, there has been little research on prosody more generally in ethnolinguistic and regional varieties of English (Burdin, Holliday, & Reed, 2018). In the present study, we address these gaps in research by investigating the extent to which listeners perceive specific aspects of intonational variation as indexes of ethnic identity.

In addition, research in perceptual sociolinguistics has rarely confronted the issue of whether social meanings are *incremental*—that is, how the social meanings of gradient

features are affected by these features' phonetic shape. Put differently, does a more phonetically extreme token of a socially marked variable correspond to a stronger social meaning? This gap is partially due to the common practice of treating continuous socially marked variables as categorical, such as /ɪ/ vocalization and /aɪ/ monophthongization (e.g., Labov, Ash, & Boberg, 2004). Even when investigating inherently continuous variables such as vowel quality, research on social meanings also tends to bin variables into discrete categories (Villarreal, 2018). In the present study, we address these gaps by investigating whether listeners' judgments of aspects of intonation are sensitive to the strength of the variable of interest in the phonetic signal.

We pursued these questions about intonational variation and social meaning via a task in which listeners rated samples of President Barack Obama's speech on the degree of 'sounding black.'¹ Critical stimuli contained either one or more L + H* pitch accents or no L + H* pitch accents. The L + H* pitch accent has been shown in production studies to be a resource for performance of African American identity (Holliday, 2016; McLarty, 2018). Pitch accents in critical stimuli also varied according to degree of phonetic extremeness (i.e., the magnitude of F0 excursions). Listeners perceived stimuli with at least one L + H* token as sounding more black than those without, but only for stimuli with more phonetically extreme L + H* realizations (i.e., those with a larger difference between F0 maximum and minimum). These findings contribute to our understanding of how listeners make ethnic judgments based on intonational variation, and how listeners assign social meaning to gradient phonetic variation.²

1.1. Ethnic identification in the U.S.

A body of linguistic research on ethnic identification dating back nearly 70 years has found that U.S. listeners are generally rather accurate (70–100%) at distinguishing black speakers from white speakers (cf. Thomas & Reaser, 2004). Recent studies have attempted to unpack the role of suprasegmentals in ethnic identification. Thomas and Reaser (2004) found that listeners were equally accurate at ethnic identification for monotonized and unmodified stimuli, suggesting that listeners do not rely solely on F0 cues in ethnic identification. They also discovered that some cues relevant to pitch accents are recoverable even from monotonized stimuli (i.e., amplitude, duration, and segmental qualities), so it is conceivable that pitch accents may aid identification even in monotonized stimuli. Holliday and Jagers (2015) examined listeners' ability to identify the ethnicity of U.S. politicians based on single-word stimuli, in order to assess the effects of voice quality on listener judgments. Building on some of the earlier findings of Purnell et al. (1999), Holliday and Jagers found that several suprasegmental variables, including jitter and harmonics-to-noise ratio, influenced ethnic identification, though they note that a combination of multiple speakers and contexts may cause challenges in isolating speaker-specific variables influencing ethnic identification. For this reason, in the present study, we attempt to control for the effect of speaker-specific voice quality variation and more carefully isolate the prosodic variables that may affect ethnic identification by employing stimuli from a single speaker.

¹ Listeners were intentionally not provided guidance on how to interpret this question, because earlier ethnic identification studies allowed for speakers to answer with their own conceptualizations of race and ethnicity (cf. Thomas & Reaser, 2004). Since one of the aims of the current study was to test for incrementality in ethnic judgments, it was important that listeners' judgments were shaped by their own state of knowledge about ethnolinguistic patterning of intonational variation.

² Portions of this data appeared in print in the University of Pennsylvania Working Papers, Selected Papers from NWAV46, as "How black does Obama sound now?: Testing listener judgments of intonation in incrementally manipulated speech."

1.2. Intonational variation: Pitch accents

This study focuses on one particular type of intonational variable as a starting point for understanding how listeners may react to ethnically-linked suprasegmental features, using methods based in the auto-segmental/metrical (AM) intonational framework (Pierrehumbert, 1980). Essential to the AM theory is the idea that movements in fundamental frequency (F0), the main correlate of what we perceive as pitch, result from an underlying sequence of tones that determine their structure. In the AM theory, these tones are either low or high, and all movements of the pitch contour are composed of a series of low and high sequences. The labeling system for intonational phenomena that is based on the AM theory is called the Tones and Breaks Index system (ToBI). Each language, and indeed a number of dialects and varieties, have distinct ToBI systems that reflect the variety's intonational specifications (Beckman & Ayers-Elam, 1997). The ToBI system for Mainstream American English (MAE), originally developed by Beckman and Ayers-Elam (1997) and based on the findings of Pierrehumbert (1980), is the only ToBI system generally in use for examining variation within American English. MAE-ToBI has previously been used for descriptions of Jewish English (Burdin, 2015), Appalachian English (Reed, 2016), as well as African American Language (AAL) (Holliday, 2016; Jun & Foreman, 1996; McLarty, 2018).³

MAE-ToBI contains two types of pitch movements: pitch accents, which occur on some stressed syllables, and edge tones, which occur at phrase boundaries. The current study focuses only on the movement of pitch accents, though it is important to note that we also tested for the perceptual effects of edge tones. This study focuses on the difference between two types of pitch accents in MAE: a simple high tone, labeled as H*, and a fall-rise, labeled as L + H*. Though other types of pitch accents exist, H* and L + H* are by far the most common pitch accents in most varieties of U.S. English, including AAL (Burdin et al., 2018).

Earlier studies have shown that pitch accents are perceptually salient for listeners and that naïve listeners can be trained to identify them quickly (McLarty, Vaughn, & Kendall, 2017; Thomas, 2011). Especially relevant to the current study, studies such as Loman (1975), Holliday (2016), and McLarty (2018) have found that MAE and AAL exhibit different rates and contexts of use for H* versus L + H*. In particular, Loman (1975) and McLarty (2018) each found that L + H* pitch accents are more common in some varieties of AAL.

Recent work by Holliday (2016), Burdin (2015), and Reed (2016) *inter alia* has also found that a greater rate of use of the L + H* pitch accent may also be a resource in production for performance of different types of ethnic identity. For example, Holliday (2016) recorded 25 men (age 18–32) with one black parent and one white parent in Washington, DC to examine their rates of use of different types of pitch accents in ethnic identity performance. The participants were recorded in casual peer dyad conversations, and the analysis of their intonational patterns was taken from these recordings. A sociolinguistic interview also elicited ideologies about race and self-identifications. The participants who identified more as black, as opposed to multiracial or mixed, were more likely to use a greater quantity of L + H* accents than H* accents. This finding supports Loman's (1975) and McLarty's (2018) findings that L + H* is more prevalent in AAL than in MAE; also relevant for the current study, this finding demonstrates that speakers' production of intonational variation is gradient in terms of frequency.

³ Though some scholars have posited that the intonational phonological inventory of AAL may differ from that of MAE, it is still considered a reliable method for analyzing intonation in AAL, at least until researchers further investigate development of an AAL ToBI system (Holliday, 2016; Thomas, 2015).

1.3. Incrementality in intonation and perception

This study's focus on intonational and suprasegmental variation presents an opportunity to address questions about phonetic detail and social meaning. One of the most significant recent advances in sociolinguistic theory has been the advent of sociophonetics (e.g., Foulkes & Docherty, 2006), with the notion that paying attention to phonetic detail can enrich our understanding of sociolinguistic variation—especially for variables that have traditionally been considered binary or categorical (Labov, Ash, & Boberg, 2006).

Although the binary treatment of phonetic variables reveals structure in sociolinguistic variation, a sociophonetically informed approach recognizes that the distribution of these variables' continuous acoustic correlates is not always compatible with discrete categorization. For example, Jacewicz and Fox (2018) use a continuous measure of /aɪ/ monophthongization (trajectory length) to analyze preadolescent Appalachian English speakers. They find that these preadolescents produce variants that are more diphthongal than Appalachian adults but less diphthongal than central Ohio adults. The authors' continuous approach pays off, in other words, by revealing finer-grained phonetic variation than is suggested by the monophthong/diphthong binary.

At the same time as research on production in sociolinguistics has increasingly turned to phonetic detail, the role of such detail remains under-theorized and under-investigated in the study of social meaning. To that end, Podesva (2011) proposes a framework for salience in sociolinguistic variation that reconciles the roles of frequency and phonetic detail. He hypothesizes that salience takes one of two linguistic forms: 'categorical salience' (frequent productions of a marked feature are salient) and 'phonetic salience' (more extreme productions are salient). In particular, with respect to phonetic salience, Podesva argues that a more extreme production signals a stronger social meaning: "If an axis of phonetic variation indexes a particular social meaning, then outliers on that axis can be understood as the *strongest indicators of meaning*" (pp. 254, emphasis added).

These predictions about categorical and phonetic salience have been supported by a handful of findings on the distribution and social meaning of intonational variation in production. For example, Podesva (2011) found that one speaker constructed a 'life of the party' persona by using acoustically extreme falling contours to imbue partying-related narrative elements with extra emphasis. Burdin et al.'s (2018) comparison of L + H* pitch accents in Jewish English, AAL, and Appalachian English showed that both categorical and continuous properties of pitch accents are sites for sociolinguistic differentiation. The authors found that, across communities, L + H* pitch accents differed in both rates of use and acoustic properties (e.g., peak F0, peak offset).

As far as we are aware, only a handful of perceptual studies have investigated how social meanings are affected by phonetic detail. Plichta and Preston (2005) presented U.S. listeners with a synthesized continuum from monophthongal to diphthongal /aɪ/ and asked listeners to identify the speaker's geographic origin along an axis running from the U.S. north to the U.S. south. Listeners not only associated monophthongal /aɪ/ with the south and diphthongal /aɪ/ with the north, they also placed successive continuum steps linearly along the north–south axis. D'Onofrio (2018) found that labeling a speaker as a 'Business Professional' or 'Valley Girl' cued U.S. listeners to classify more ambiguous [æ~ɑ] tokens as /æ/, with 'Valley Girl' being especially associated with backer /æ/, compared to a 'Chicago Bears Fan' label or no label at all. In an experiment with Californian listeners, Villarreal (2016) found significant correlations between speakers' raising of /æ/ in *bad* and *glass* and listeners' ratings on the scales 'accented,' 'doesn't speak like me,' 'unfamiliar,' and 'not Californian.' Foulkes, Docherty, Khatlab, and Yaeger-Dror (2010) found that listeners' identification of Tyneside children's gender was affected by two continuous measures (amplitude and F0) as well as several categorical measures; however, the authors also

report significant correlations between amplitude and F0 in stimuli, suggesting potential issues with collinearity in the modeling procedure. In terms of voice quality, Szakay (2012) found that in New Zealand, ethnic identification was affected by several continuous voice quality measures; speakers with higher mean H1–H2 (a measure of creakiness) were likelier to be identified as Māori.

The present study seeks to expand our understanding of the relationship between phonetic detail and social meaning by investigating this relationship through the lens of intonational variation. Building on Podesva (2011), we hypothesize that the social meanings of continuous variables will exhibit what we call *incrementality*: a monotonic relationship between the variable's phonetic extremeness and the strength of the social meaning it elicits in perceivers.⁴ We focus on pitch accents, which are ideally suited to this question as they vary both in category (e.g., H* versus L + H*) and phonetic shape (e.g., peak offset, rise slope).

2. Methods

This study was designed to address three central research questions:

1. How do pitch accents affect listener judgments of ethnic identity? In particular, does the L + H* pitch accent carry a social meaning of blackness in perception, as it does in production?
2. To what extent are the ethnicity-based social meanings of these pitch accents mediated by incremental phonetic differences?
3. What other aspects of voice quality affect listener judgments of ethnicity?

These questions were investigated via a perceptual task in which listeners rated 120 samples of President Barack Obama's speech with respect to how much they thought he 'sounded black' in each particular sample.

2.1. Open-guise versus matched-guise technique

This task used the 'open-guise technique' (OGT) (Soukup, 2013); as in the more common matched-guise technique (MGT), OGTs offer insight into the social meanings of a focal feature, variety, or language, by comparing listeners' reactions to stimuli differing only by the focal linguistic structure (e.g., Campbell-Kibler, 2009). Unlike the OGT, the MGT axiomatically hinges on listeners' belief that they are listening to different speakers (Giles & Billings, 2004; Purnell et al., 1999); otherwise, it is assumed that listeners will not differentiate guises on personal characteristics that are considered intrapersonally stable qualities (e.g., intelligence). In OGTs, by contrast, listeners are openly informed that they are hearing the same speaker in different guises. Soukup (2013) shows that listeners responded differently to standard versus dialectal Austrian German guises in both OGT and MGT settings (with the OGT actually yielding stronger effects for some scales), undermining MGTs' key assumption about different speakers.

In the present study, we assumed that listeners (all from the United States) were highly likely to recognize our stimulus speaker, President Barack Obama, necessitating an OGT rather than MGT approach. We openly informed our listeners, "This study is designed to test how people respond to different speech excerpts from the same speaker." In so doing, we rejected the type of instrumental task framing often used in MGTs, such as evaluating prospective radio newsreaders (Labov et al., 2011; Villarreal, 2018). By contrast, Obama

⁴ Whereas Podesva's phonetic salience hypothesis applies only to phonetic outliers, our incrementality hypothesis applies across the 'axis of phonetic variation'; the latter can thus be considered a stronger form of the former.

represented an ideal stimulus speaker to test our hypotheses, as his ability to command both AAL and MAE is well-known by the general public (Alim & Smitherman, 2012); our use of the OGT took advantage of this awareness. In the discussion, we make recommendations about the appropriateness of OGT versus MGT.

2.2. Stimulus creation

The 120 stimuli were based on excerpts of President Barack Obama's spontaneous speech from two different 2016 television interviews with Gayle King, a black broadcast journalist who co-anchors the *CBS This Morning* news program (Kaplan, 2016). Each stimulus excerpt was based on a single Intonational Phrase (IP) unit, ranging from 0.4 to 2.3 seconds in duration (median 0.9 seconds). Following Pierrehumbert and Hirschberg (1990) as well as subsequent works utilizing their methods, we identified IPs through looking for pausing and phrase-final lengthening, as well as the presence of characteristic boundary tones and smaller intermediate phrase units contained within the IPs. We attempted to select short phrases that were fairly semantically bland to avoid overly tilting responses in one direction, though it is impossible to completely control for content in listening tasks.

Sixty excerpts were selected: 20 critical excerpts and 40 filler excerpts. Ten critical excerpts were 'H* phrases,' which contained between 1–3 H* pitch accents and 0 L+H* accents; and ten were 'L+H* phrases,' which contained between 1–3 L+H* pitch accents and 0–2 H* accents. This imbalanced definition of H* versus L+H* phrases was necessary since L+H* pitch accents are relatively rarer, even in AAL (Burdin, Holliday, & Reed, 2018), so it was not possible to find enough excerpts that contained only L+H* accents. Filler excerpts contained 1–3 H* pitch accents and 0 L+H* accents.

In choosing excerpts, we intentionally sacrificed a degree of experimental control for the sake of presenting listeners with natural, spontaneously produced stimuli rather than unnatural, lab-like speech. The benefit of using spontaneous stimuli is that it more closely models real-world perception conditions, as listeners perceive spontaneous and read speech (including oratory) differently (Campbell-Kibler, 2009; Holliday & Jagers, 2015). The drawback is that the distribution of H* versus L+H* pitch accents across stimuli prevented us from addressing Podesva's (2011) hypothesis about categorial salience; at the same time, total experimental control over stimuli is impossible to obtain, as features co-occurring in the stimuli can always shape interpretation of features of interest (Leach, Watson, & Gnevsheva, 2016), including propositional content (Campbell-Kibler, 2009). (We explore this issue further in a post hoc analysis of L+H* phrases.)

The critical stimuli were created by manipulating critical excerpts to four manipulation steps, with the original excerpt as Step 1. Steps 2, 3, and 4 were created by making pitch accents' F0 minima and maxima successively more extreme. With each manipulation step, H* and L+H* maxima were increased by a semitone, and L+H* minima were decreased by a half-semitone. For example, the H* pitch accent in the top panel of **Figure 1** has an F0 maximum at 118.3 Hz in step 2 and 125.2 Hz in step 3, a one-semitone difference; the L+H* pitch accent in the bottom panel has an F0 minimum at 101.6 Hz in step 1 and 99.1 Hz, a half-semitone difference. (In some cases, it was not possible to make the manipulations exactly one or one-half semitone.) We based F0 manipulations on semitones rather than constant magnitudes because semitones are psychoacoustically comparable regardless of the pitch accent's initial F0 (e.g., the difference between 100 and 105 Hz sounds much larger than the difference between 200 and 205 Hz). The first author created stimuli by hand using the Manipulation utility in Praat (Boersma & Weenink, 2015). Both authors listened to all manipulated critical stimuli and confirmed that they sounded natural.

Filler stimuli were created by modifying the final syllable of filler excerpts to include percepts of creaky voice: low F0 and damped pulses (Keating, Garellek, & Kreiman,

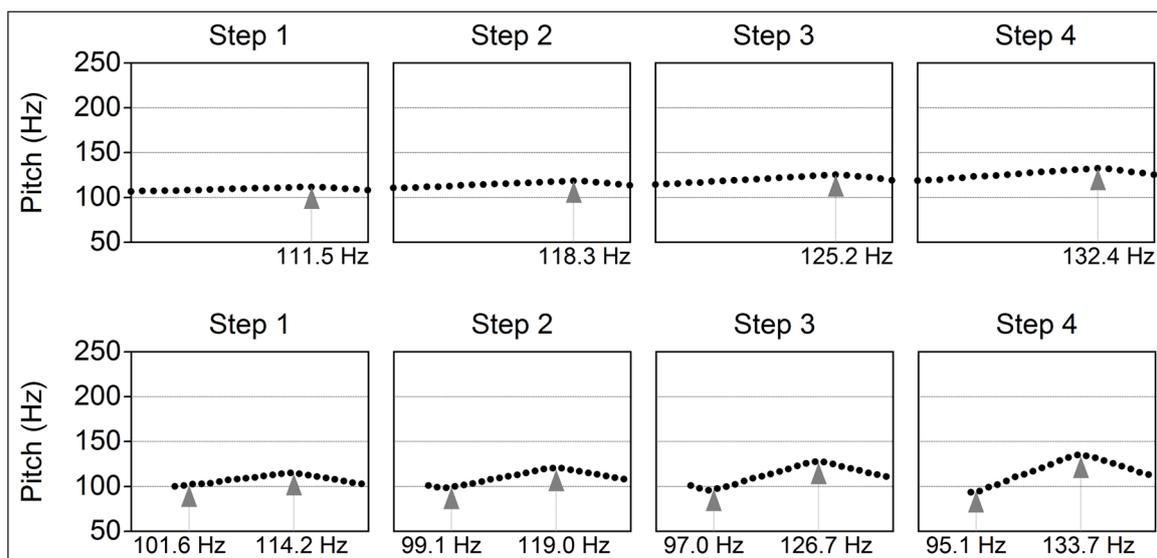


Figure 1: Original (Step 1) and manipulated (Steps 2–4) versions of pitch accents in stimuli: H* pitch accent in *would* (top) and L+H* pitch accent in *all* (bottom).

2015). A Praat script modified alternating cycles of the final syllable by lengthening their duration and lowering their amplitude. As with critical stimuli, both authors listened to all manipulated filler stimuli and confirmed that they sounded natural.

2.3. Task design

The task was administered via an online survey hosted by Qualtrics. In each of 120 randomly ordered trials, listeners heard a single stimulus auto-play twice and responded to the question “How black or white does Obama sound here?” on a continuous unit-less slider bar with “very black” and “very white” on opposite poles. As the recognizability of President Obama’s voice would have likely rendered ineffective the type of instrumental task framing often used in MGTs (e.g., rating prospective radio newsreaders, as in Labov et al., 2011), we eschewed such framing; we instead informed listeners, “This study is designed to test how people respond to different speech excerpts from the same speaker.” Listeners then completed a demographic questionnaire and were invited to comment on the task (see Appendix A).

The survey was distributed via social network sampling in May 2017, with a raffle incentive for one randomly selected listener to win an Amazon.com gift card. The listener sample for analysis contains 79 American English-speaking listeners who self-identified as black and/or white. Of these listeners, 24% self-identified as black and 77% as white (one listener identified as both); 65% identified as female and 35% as male. The majority of listeners also self-identified as politically liberal and indicated that they overwhelmingly approved of Obama’s presidency; in particular, on a 1–7 scale (where 7 indicated “very liberal” and “strongly approve of Obama”), the median rating was 6 on both scales, and 91% of listeners rated 5 or above on *both* scales. In this respect, the listener sample is not representative of the United States voting population; however, our intent was not to survey a sample spanning the political spectrum but rather to determine how some listeners judge ethnicity based on intonational and voice quality variation (we return to this point in the Discussion).

As mentioned above, both authors listened to all stimuli and confirmed that they sounded natural. As a further check on stimulus naturalness, we coded listeners’ responses to the final two questionnaire items: “How did the clips sound to you?” and “Do you have any other comments on the clips or on the survey?” Based on listeners’ responses

to these questions, the second author developed eight true-or-false codes that described sentiments listeners expressed in their responses and coded responses accordingly (with a single response capable of being coded “true” in multiple categories). For example, 21% of listeners reported something amiss with the quality of the clips (although numerous listeners commented positively about the clips’ quality). More information about these codes, including examples, can be found in Appendix B. As we discuss below, however, none of these codes significantly improved our model of intonation results, so we did not find evidence that they impacted listeners’ perceptions of the speaker’s blackness.

Slider-bar positions were converted to real numbers between 0 (“very white”) and 100 (“very black”) and standardized by listener to control for variable usage of the continuous slider bar. All results are reported in unit-less standard deviations (i.e., z-scores); the average listener’s standard deviation was 16.6, so a difference of 1 standard deviation can be interpreted as a difference of roughly one-sixth of the length of the slider bar for the average listener.

Our task was specifically designed to address the first two research questions, about the role of pitch accents and phonetic incrementality in affecting listener judgments of ethnicity; we first present the analysis of intonation features. We then describe a post hoc analysis of voice quality characteristics that addressed the third research question, about the role of other voice quality features in affecting listener judgments of ethnicity.

3. Intonation analysis

We compared linear mixed-effects models of standardized ratings to find the predictor structure that best modeled the data in critical trials, via the `lmerTest` package for R (Kuznetsova, Brockhoff, & Christensen, 2016; R Core Team, 2018). The predictors that we tested were phrase type (H* versus L + H*), manipulation step, edge tone, nuclear pitch accent, stimulus duration, and numerous listener effects (race, gender, political ideology, approval of Obama’s presidency, education, use of desktop versus mobile to complete survey, hometown, geographic mobility, experience with linguistics, musical experience, and qualitative questionnaire codes). Unfortunately, the distribution of H* and L + H* tokens in stimuli precluded predictors for the number of H* and number of L + H* pitch accents in critical trials. We also included random intercepts for excerpts as nested within phrase type, as each excerpt exclusively belonged to one of the two phrase types. Since ratings were standardized by listener, by-listener random intercepts would be redundant.

Table 1 presents a summary of the best model for listener rates of blackness, which included predictors of phrase type, manipulation step, and their interactions.

Table 1: Summary of best model of listener ratings of blackness. Degrees of freedom estimated via Satterthwaite approximations (Satterthwaite, 1946). Significance: * $p < 0.05$.

	Estimate	SE	d.f.	t	p
(Intercept)	-0.0409	0.1065	23.1	-0.384	0.7045
PhrTypeL+H*	0.0172	0.1507	23.1	0.114	0.9103
Step2	0.0041	0.0458	6056	0.089	0.929
Step3	-0.0214	0.0459	6056	-0.466	0.6415
Step4	0.0132	0.0459	6056	0.287	0.7737
PhrTypeL+H*:Step2	0.0297	0.065	6056	0.457	0.6475
PhrTypeL+H*:Step3	0.1314	0.065	6056	2.02	0.0434 *
PhrTypeL+H*:Step4	0.1047	0.065	6056	1.609	0.1076

As is evident from this model, listener ratings of blackness tended to increase with the more extreme step manipulations, though this is only statistically significant for L + H* phrases. Also notable is that the model revealed no significant listener effects for gender, race, region, education, or political affiliation, indicating that listeners were remarkably similar in their ratings regardless of a number of potentially influential demographic factors. While previous studies have generally found that in-group community members may perform better in ethnic identification tasks (cf. Thomas & Reaser, 2004), there were no such effects observed here. In addition, none of the qualitative questionnaire codes significantly improved the model; this means that, for example, although some listeners commented negatively on the quality of the stimuli, we have no evidence that whether or not listeners commented on stimulus quality affected listener perceptions of blackness.

These results must be interpreted with caution, however, in light of their small effect size. The sole significant term in **Table 1**, PhrTypeL + H*:Step3 (which has a *p* value just under the predetermined α level of 0.05), differs from the intercept by about 0.17 standard deviations, or about 3 ‘notches’ on the 0–100 slider bar (with the average listener’s standard deviation being 16.6). Indeed, an R^2 calculation using the R package piecewiseSEM (Lefcheck, 2016) revealed that the model’s fixed-effects predictor structure accounted for less than 1% of the variance in ratings, while random effects—the effect of individual excerpts—accounted for 11.3% of the variance.⁵ With that caveat in mind, we proceed to discuss what these results mean.

3.1. Results by phrase type

The model indicated no main effect of phrase type on listener ratings of blackness, indicating that pitch accent alone did not trigger different blackness ratings. **Figure 2** shows this result, with results for H* stimuli in the left panel and L + H* stimuli in the right panel.

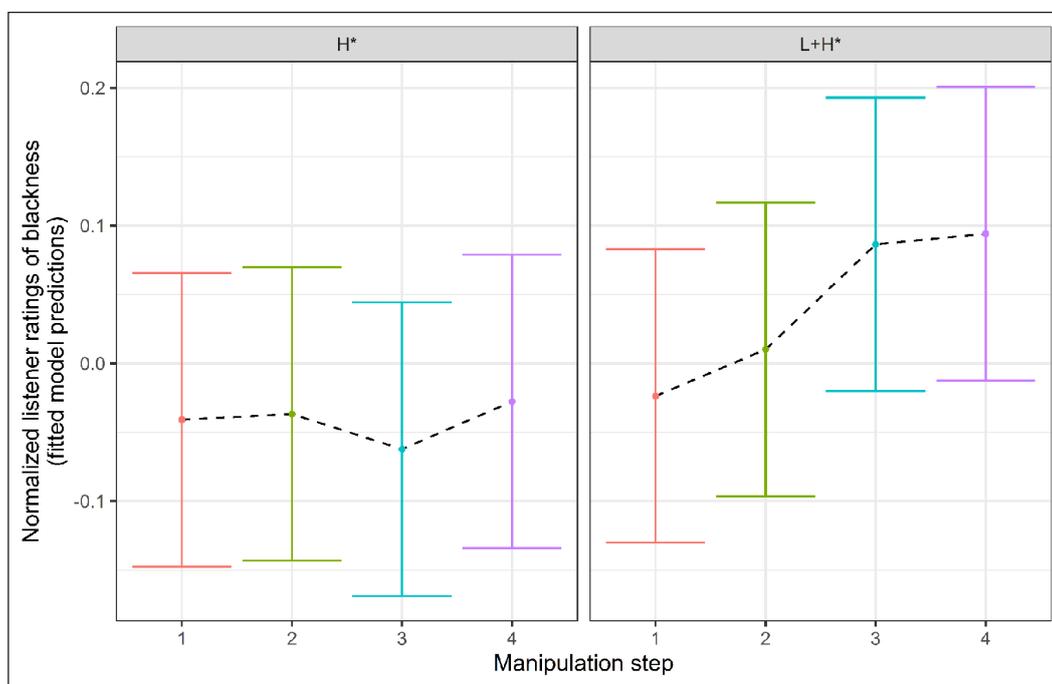


Figure 2: Fitted model predictions for listener ratings of blackness by phrase type and manipulation step. Error bars represent 95% confidence intervals.

⁵ An anonymous reviewer expresses doubt that this significant effect “would reliably reappear as an important factor” in a replication of the present study; we agree that this is an empirical question.

As is evident in this figure, listener ratings of blackness were remarkably similar for the H* phrases at each step, though the L + H* phrases showed greater differences between manipulation step. There was also no main effect of manipulation step on listener ratings of blackness.

3.2. Results by manipulation step

Though the main effect of phrase type failed to reach significance, the model indicated a significant interaction between phrase type and manipulation step, with more extreme L + H* phrases rated as sounding blacker than less extreme L + H* phrases, and no perceived blackness difference for H* phrases regardless of step. **Figure 3** presents these results, with each panel representing a manipulation step. This figure indicates that listeners appear to interpret the more phonetically extreme L + H* realizations (greater difference between F0 minimum and F0 maximum within a L + H* pitch accent) as blacker, but this is not the case for the more extreme H* realizations (which only had higher F0 maxima).

This model also implies that listener judgments of blackness are affected by more than just pitch accent type and phonetic shape. As mentioned above, these results must be interpreted with caution, especially in light of the fact that the model’s fixed-effects predictor structure accounted for less than 1% of the variance in ratings, while random effects—the effect of individual excerpts—accounted for 11.3% of the variance. In other words, listeners were much more attuned to features varying by excerpt, such as segmental, semantic, pragmatic, or voice quality characteristics, than the type and phonetic shape of pitch accents. However, this small effect size may represent an inherent challenge to studies of prosody, since the highly nested nature of such variables causes them to be difficult to isolate from one another. Despite this challenge, the finding of a significant

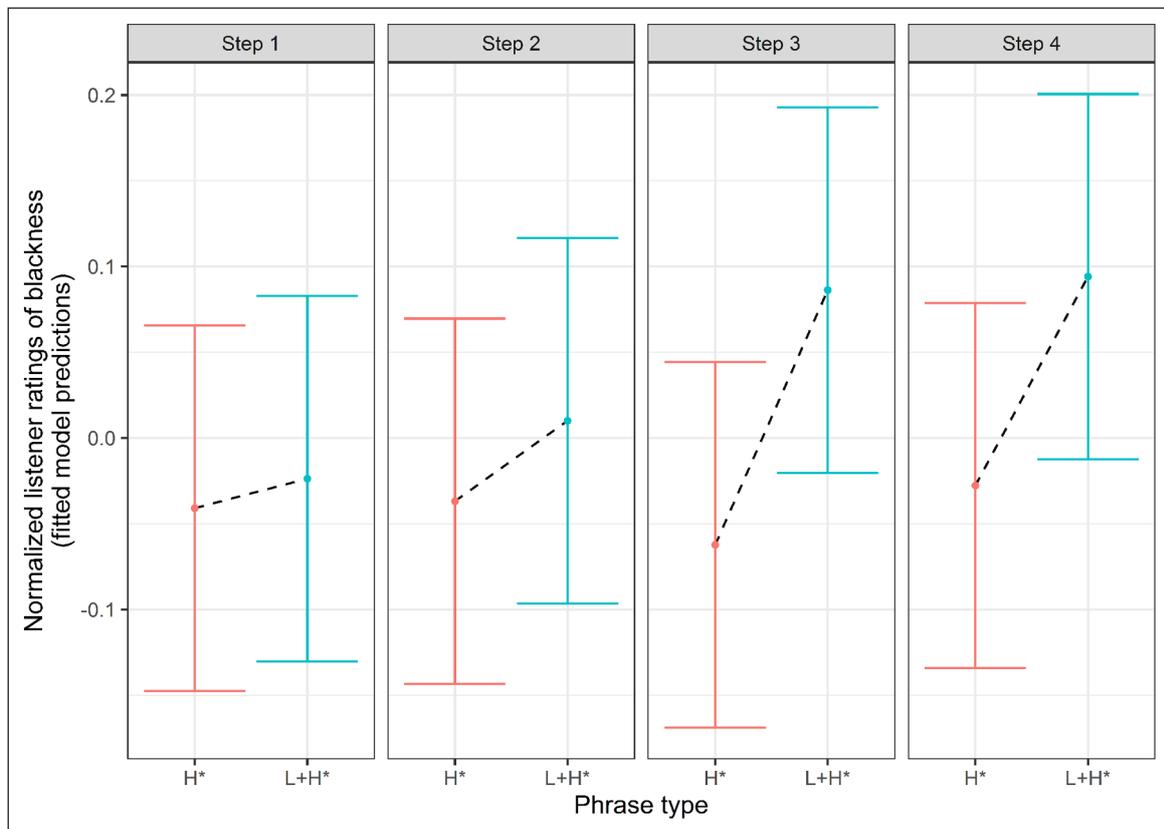


Figure 3: Fitted model predictions for listener ratings of blackness by manipulation step and phrase type. Error bars represent 95% confidence intervals.

difference here may be a step in the direction of discovering how these variables may operate both independently and together. The small effect size of this intonation effect motivated the post hoc analysis of voice quality features.

4. Voice quality analysis

Our perceptual experiment was specifically designed to test predictions about how listener judgments of ethnicity are influenced by the type and phonetic shape of pitch accents; however, sociophoneticians have long suspected that voice quality characteristics may also influence listener judgments of ethnicity (e.g., Holliday & Jagers, 2015; Purnell et al., 1999). In line with Purnell et al. (1999), we conducted a post hoc analysis of perceived blackness ratings to determine if and how a number of voice quality measures were influential in shaping listener judgments. In particular, the results of their study indicate dialect-level differences in both harmonics to noise ratio (HNR) and peak pitch ratio, so we hypothesized that these same variables may also be of interest in the current study.

We ran a Praat script on critical stimuli to extract several measures that, according to previous studies, may pattern differently in AAL versus MAE: phrase speech rate, pitch ratio (Holliday & Jagers, 2015), peak delay (Holliday, 2016; Reed, 2016), jitter (Holliday & Jagers, 2015), shimmer (ibid.), HNR (Purnell et al., 1999), and intensity average (ibid.).⁶ Phrase speech rate was calculated as the stimulus's duration divided by the number of syllables. Pitch ratio was calculated as the stimulus's maximum F0 (in Hz) divided by its minimum F0. The remaining measures were calculated for each pitch accent in each stimulus; since listeners reacted not to individual PAs but whole stimuli, for stimuli with multiple PAs we treated the mean of each PA's measurement as the measurement for that stimulus (e.g., we defined the jitter measurement for a stimulus with three PAs as the mean of the PAs' jitter measurements). Peak delay was calculated as the time difference between nucleus onset and hand-annotated pitch accent time. Jitter (relative average perturbation), shimmer (local amplitude perturbation), HNR, and intensity average (mean dB) were all calculated for the nucleus. An F0 floor of 75 Hz was used for all relevant measures in order to avoid erroneous measurements of non-periodic speech; we otherwise used Praat's default settings for all measurement functions.

As with the intonation analysis, we modeled standardized ratings via linear mixed-effects models. Because the intonation analysis revealed differences in patterning of responses to H* versus L + H* stimuli, we fit separate models to H* versus L + H* critical trials. We included manipulation step in these models to determine whether the intonation analysis's findings about the role of manipulation step—significantly affecting listener ratings of blackness in L + H* stimuli but not H* stimuli—remained after considering voice quality features. These models also included random intercepts for excerpts and random by-excerpt slopes for the manipulation step factor. Voice quality measures were normalized (z-scored) to account for widely differing measurement scales.

To account for likely collinearity of voice quality measures (e.g., jitter and pitch ratio are all different measures of changes in fundamental frequency), we adopted a model-comparison strategy that iteratively added interaction terms to the models based on correlations between measures. We first ran baseline models that included all voice quality measures as main effect predictors with no interactions. (Again, these models also included random intercepts for excerpts and random by-excerpt slopes for the manipulation step factor.) We then checked these baseline models for correlations between voice quality measures; any correlations with an absolute value correlation coefficient greater than 0.4 in either model

⁶ Phrase speech rate, peak delay, and vowel duration are prosodic features, not voice quality features, but for the sake of brevity we refer to the entire set as voice quality features.

were added as interaction terms into both models. After running these models, we again added interaction terms (including three-way interactions) based on correlations between voice quality terms. The resulting models included the following interactions: phrase speech rate \times peak delay \times HNR, shimmer \times jitter \times HNR, pitch ratio \times intensity average. For both the H* and L + H* models, each successive model represented a significant improvement in model fit at an $\alpha = .05$ significance threshold.

4.1. Voice quality results

Summaries of fixed effects for the voice quality models are in Appendix C. The voice quality model for H* critical trials revealed that few voice quality measures significantly affected listener perceptions of blackness: phrase speech rate and the interaction of peak delay and HNR. Phrase speech rate (seconds per syllable) had a positive effect on listener perceptions of blackness, with slower phrases rated blacker. While the model returned positive estimates for the effects of peak delay and HNR, neither of these main effects reached significance. Rather, the effect of peak delay on listener perceptions of blackness was constrained by the phrase's HNR. As **Figure 4** shows, phrases with longer peak delay were rated blacker, but only if HNR was sufficiently high. This co-patterning of variables suggests that listeners may be attuning to a threshold of combined characteristics in order to make judgments, particularly in the absence of ethnolinguistically-salient intonational differences such as the L + H* pitch accent. Together, the roles of peak delay and phrase speech rate suggest a possible salience effect, as both relate to vowel duration; conceivably, longer vowels (which co-pattern with slower speech rates) with longer intonation rises can better carry indexes of social meaning.⁷

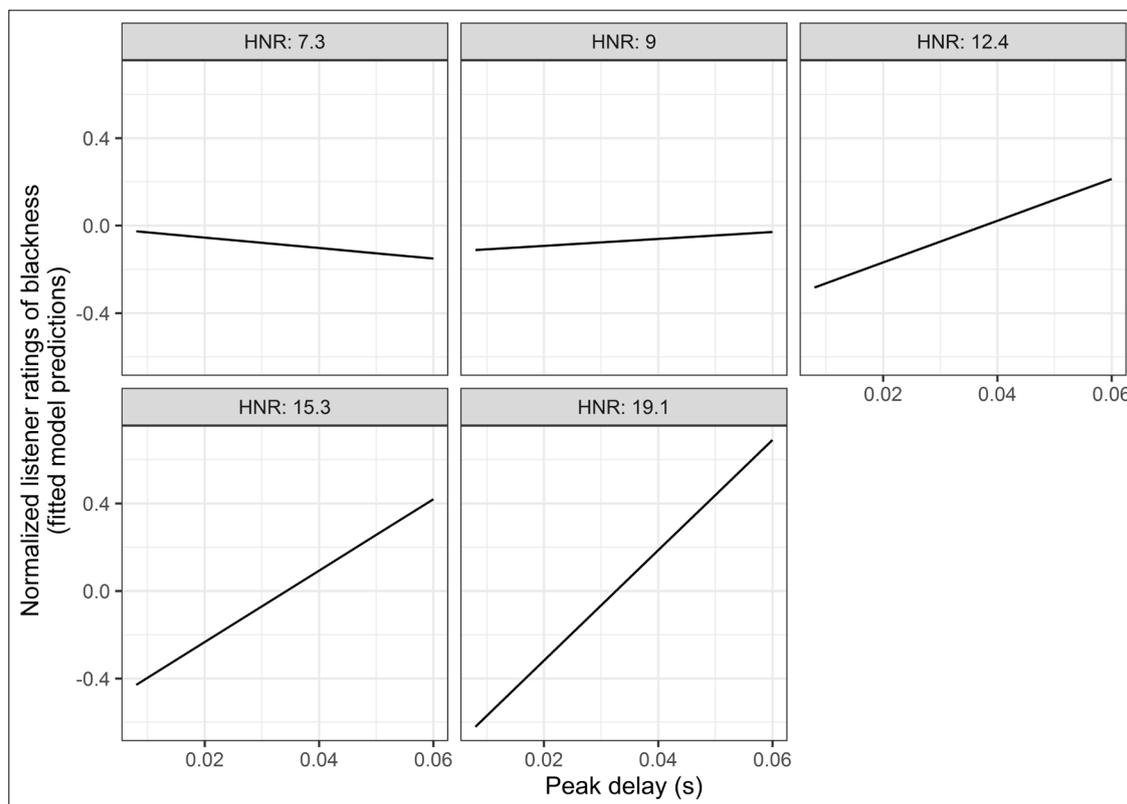


Figure 4: H* model predictions for perceived blackness ratings by peak delay (seconds) and HNR (dB). The five facets display peak delay slopes at the minimum, first quartile, median, third quartile, and maximum values for HNR among H* stimuli.

⁷ Thanks to an editor for pointing this out.

As with the H* model, few predictors reached significance in the L + H* model—including just one voice quality measure, jitter. Among L + H* stimuli, phrases with less jitter were rated blacker, suggesting that listeners are sensitive to the interaction of F0 movement and local periodic perturbations. Notably, the measures affecting listener perceptions of blackness did not overlap for H* versus L + H* phrases; the jitter term in the H* model, and the phrase speech rate & peak delay × HNR terms in the L + H* model, did not even approach significance. This finding provides additional evidence that listeners may respond to different intonation and voice quality cues in phrases containing L + H* pitch accents than those not containing L + H* pitch accents. As L + H* accents are far less common than H* accents, it is possible that L + H* accents cue listeners to adjust their expectations as to markers of ethnic identification.

In addition, manipulation step was significant in the L + H* voice quality model (manipulation steps 3 and 4 were rated blacker than steps 1 and 2) but not the H* voice quality model. This finding corroborates the generalization that the percept of blackness is subject to phonetic incrementality only with respect to the more socially marked L + H* pitch accent. However, this finding is tempered by the fact that the fixed effects in the L + H* model accounted for less than 3% of the variance, as compared with 10.2% for the fixed effects in the H* model (Table 2); in other words, there remain properties of the stimuli with L + H* accents that listeners are reacting to, above and beyond the intonational and voice quality features that previous studies have suggested are implicated in differentiating AAL from MAE.

In short, the voice quality analysis found that listeners relied on multiple acoustic cues—beyond those pertaining to pitch accents’ type or phonetic shape—in making judgments of perceived blackness; crucially, in the presence of an L + H* pitch accent listeners not only relied on different voice quality cues than in the absence of one, but they apparently relied to a much greater degree on cues other than those relating to intonation or voice quality. This finding suggests a fundamental difference in how listeners judge phrases in the presence of an L + H* pitch accent, although this is an open question for future study. More broadly, this finding further supports the claim that understanding the interrelated nature of prosodic variables is a necessary part of their description.

5. Discussion

To summarize, this study has demonstrated that listeners are sensitive to the details of phonetic realizations of the H* and L + H* pitch accents in declaratives, and that a larger difference between the F0 maximum and minimum within L + H* pitch accents appears to cause listeners to rate a speaker (in this case, President Barack Obama) as sounding blacker. However, the difference between H* and L + H* pitch accent phrases alone is not sufficient to trigger this judgment; it is the actual realization of the pitch accents themselves that listeners seem to attune to. In addition to pitch accent type and phonetic shape, listeners also attend to voice quality cues in judging blackness, though the relevant cues are different for H* versus L + H* phrases: speech rate, peak delay, and harmonics to noise ratio for H* phrases, jitter for L + H* phrases. There is also some evidence that the number of L + H* and H* pitch accents in a phrase also affect listener judgments of blackness.

Table 2: R² values (percentage of variance accounted for) for voice quality models, calculated via R package piecewiseSEM (Lefcheck, 2016).

	Fixed-effects R ²	Random-effects R ²	Total R ²
H* model	10.2%	4.3%	14.5%
L+H* model	2.7%	24.7%	27.3%

We also obtained an unexpected finding with respect to speech rate; among H* stimuli, slower phrases were perceived blacker than faster phrases, which could possibly indicate that speakers have different expectations related to ethnolinguistic variation and speech rate (Kendall, 2013) or that longer vowels provide a greater site for the apprehension of social meaning. In this section we discuss the implications of these findings in more depth.

5.1. Intonation

This study's results show that in a perception task, listeners appear to be sensitive not only to the phonological category of pitch accents, but also their phonetic realization, as listeners appear to be sensitive to increasingly extreme manipulations of F0 within a single pitch accent type. In the traditional AM model of intonational phonology, pitch accent and edge tones have largely been binned into discrete categories, with meaning presumed to be attached to those categories and their combinations (Pierrehumbert & Hirschberg, 1990). The results presented here provide further motivation for considering intonational variation on a phonetic as well as a phonological level. This study also provides further motivation for the development of ethnolinguistic variety-specific ToBI models as well as phonetic methods for studying intonational variation cross-dialectally. While we have employed the MAE-ToBI conventions (Beckman & Ayers-Elam, 1997) in this study, the nature of the intonational system of AAL has not yet been fully described (McLarty, 2018; Thomas, 2015). As the current study's results provide evidence that listeners are sensitive to differences in the realization of F0 and timing of the L + H* pitch accent, future studies should examine whether the tonal inventory of AAL differs from MAE, as this could be one element that triggers the observed differences in listener judgments.

Relatedly, as much of the work on prosody has focused on the meaning of intonational contours in an imagined Standard American English as opposed to in specific varieties, it is clear that much more work is needed on both variation in speaker production and listener perception of contour meaning. Though the current study did not reveal differences in perception of 'sounding black' conditioned by listener demographics, future work should explore how such perceptions could potentially be affected by listeners with different backgrounds and sociolinguistic experiences.

This point about the role of demographics is especially relevant because (as mentioned above) the listener sample was overwhelmingly liberal and approving of Obama's presidency, more so than the US population at large. While this is not an issue for the present study—our aim was not to achieve political representativeness but rather to ascertain how intonational variation affected perceptions of blackness within a population of US listeners—it does contextualize the results. Theoretical frameworks that take as primary the role of experience in forming linguistic representations (e.g., Exemplar Theory, Pierrehumbert, 2016) would take the standpoint that listeners who are more inclined to listen to President Obama would have a greater opportunity to hear him in multiple situations, thus facilitating their awareness of his style-shifting; it is thus conceivable that the small-sized effects uncovered in this study would not reach significance in a sample more representative of the US political spectrum. This is a question open for future work to address.

5.2. Ethnic identification

In their 2004 study and summary of the body of research on ethnic identification of white and black speakers and the U.S., Thomas and Reaser reveal gaps in our knowledge about what triggers judgments of speakers as 'black' or 'white.' Most ethnic identification studies have focused on segmental features, at least in part due to the fact that so little is known about how non-standard varieties of American English employ intonational variation,

though it is the case that such studies on prosodic variables have been carried out outside the U.S. (Szakay, 2012; Todd, 2002). While a number of segmental features, such as vowel quality, have been identified as important in triggering listener judgments, researchers still know relatively little about how suprasegmental features may contribute to these judgments. Dating back to the 1970s, researchers such as Tarone (1973) and Loman (1975) have suspected that suprasegmental features played a serious role in triggering these judgments, though few studies have been able to isolate the specific intonational and suprasegmental features involved. The results of the current study, especially those related to the fact that speakers are able to provide consistent judgments of how a speaker whose race is known to them adheres to their ideologies about what it means to ‘sound black,’ provide evidence that it may be possible to isolate the variables of interest using a single-speaker model. This has the advantage of eliminating other types of variation that are inherent in studies with multiple speakers, for whom it is impossible to control every level of linguistic variation, which may be important especially in light of our findings on the effects of voice quality.

This study also builds on the findings of Purnell et al. (1999) as well as Thomas and Reaser (2004), and Holliday and Jagers (2015) by providing further evidence that a number of voice quality features, including jitter, HNR, and speech rate may be involved in triggering ethnicity judgments. The pattern that we observed wherein there appear to be important interactions of intonational and voice quality features obviates the need for more controlled studies that simultaneously focus on a number of suprasegmental features. Listeners appear not only to be sensitive to both intonational and voice quality features but also the ways in which they combine to create sociolinguistic meaning.

It is worth reiterating here the small effect size that we found in our intonation model, in which fixed effects accounted for just 1% of the variance in listener ratings of blackness. Some readers may interpret this small effect size and the proximity of the sole significant intonational model term’s p value (0.0434) to our predetermined α level (0.05) as casting doubt upon the generality of the result. Although this effect size is modest, it is not without precedent in studies of sociolinguistic perception. Clopper (2010, p. 212), describing Clopper and Pisoni’s (2007) study of free classification of regional dialects of American English, notes “grouping accuracy was still rather poor overall, which may indicate attention to talker-specific differences instead of dialect-specific variation.” This greater attention to talker-specific differences parallels our finding that random effects accounted for a much greater percentage of the intonation model’s variance. Likewise, Villarreal (2018) found that out of 12 ratings scales, a vocalic guise manipulation yielded only three significant differences, compared to eight significant differences for both speaker region and speaker gender and eleven significant differences for speaker ethnicity. In other words, while the effect revealed by the intonation model is modest, it is possible that this is a general property of phonetic guise manipulations, as well as an artifact of the interconnected nature of suprasegmental features in general and the resulting challenges in isolating them from one another.

5.3. Incrementality

These findings support the notion that listeners attend to phonetic detail in constructing social meanings of sociophonetic variation, given that listener ratings of blackness for L+H* increased stepwise as L+H* pitch accents became more phonetically extreme. In other words, there is some evidence that listeners map continuous social meanings to continuous variation, supporting our incrementality hypothesis; contra Podesva’s (2011) phonetic salience hypothesis, these findings suggest that greater social meanings are not only attached to phonetic outliers, but also to phonetically intermediate realizations of

L + H* pitch accents. This research also sheds light on how phonetic salience works in context. While the intonation analysis found a jump between manipulation steps 2 and 3 in listener ratings of blackness for L + H* phrases, the analysis of the L + H* voice quality model's random effects found considerable differences in step 1 ratings across L + H* phrases. That is, for some stimuli smaller differences in intonation were sufficient to trigger higher listener ratings of blackness; for others listener ratings of blackness only increased with larger differences in intonation. Thus, just as context shapes the social meaning of a variant's presence or absence (Campbell-Kibler, 2009; Gumperz, 1982; Leach et al., 2016; Pharao, Maegaard, Møller, & Kristiansen, 2014), context also shapes the way that phonetic detail affects social meanings.

5.4. Open-guise versus matched-guise technique

These findings expand our understanding of methods for probing language attitudes, countering the received wisdom in MGT research that these tasks only work if listeners believe they are judging different speakers (Giles & Billings, 2004). This work expands on the findings of Soukup (2013) in demonstrating additional support for the OGT: Listeners were aware that they were hearing the same speaker, but the guise manipulation nevertheless yielded a difference in listener responses. Soukup (2013) finds that the OGT yielded larger effects than the MGT on 'superiority' scales (Zahn & Hopper, 1985), while the MGT yielded larger effects on 'social attractiveness' scales. However, her comparison did not address socio-indexical traits like ethnicity that fall outside the superiority-versus-social-attractiveness rubric, but which nevertheless form an important part of listeners' awareness of language variation (e.g., Hay & Drager, 2010; Koops, Gentry, & Pantos, 2008; Niedzielski, 1999). Although it is impossible to determine how the results of this study would compare to a hypothetical companion MGT (as the MGT simply wouldn't work with such a recognizable stimulus speaker)—and the small effect size we found suggests that a hypothetical companion MGT could yield larger effects—the present study indicates that a socio-indexical trait, ethnicity, *can* also work in an OGT context.

Moreover, whereas stimulus speakers in typical MGTs are anonymous to listeners, representing blank attitudinal canvases save for small bits of contextual information provided via stimulus text and/or explicit labels, listeners in this study likely had salient prior impressions of President Obama and his racialized speech. The finding that the guise manipulation affected listener perceptions of Obama's blackness is even *more* persuasive against that backdrop. Indeed, among the qualitative questionnaire codes that failed to significantly improve the model was *ObamalsBlack* (see Appendix B); that is, we found no evidence that listener perceptions of blackness were affected by whether listeners found it difficult to rate Obama as 'sounding white.'

Although the OGT worked in the present study, we caution readers against the assumption that the OGT will necessarily apply to any context, feature, or trait. First, while both Soukup's study and the present study intentionally violated the assumption that listeners should believe they are judging different speakers, in both studies listeners were not told which *feature* was manipulated; we argue that this remains an important element of methodological opacity in speaker evaluation tasks. It is likely that doing so would produce rather different results than if listeners are not informed, especially for those few sociolinguistic variables that attract public commentary. Indeed, only 20% of the listeners in the current study reported that they could detect the guise manipulation (DetectManip, Appendix B), and this failed to significantly improve the model; this is helped by the fact that, aside from high rising terminal (Tyler, 2015), intonational variation is generally not a subject of public commentary in American English.

Second, we argue that there remain contexts in which it is important to conceal the fact that the same speaker is behind both or all guises. While the majority of speaker evaluation tasks involve cognitive and/or affective responses, we predict that tasks involving behavioral responses (e.g., making a hiring decision) are likelier to hinge on listeners believing they are hearing different speakers. For example, if the landlords in Purnell et al. (1999) knew they were hearing John Baugh in multiple guises, they might have been on their ‘best behavior’ to avoid prosecution under the Fair Housing Act.

Third, we argue that the use of an OGT rather than MGT approach must be justified by a plausible style-shifting context. For example, this task relied on listeners’ awareness of President Obama’s style-shifting to sound more black in some contexts and less black in others (Alim & Smitherman, 2012); as mentioned above, it is conceivable that listeners’ awareness in this respect was facilitated by their generally positive attitude toward Obama’s presidency making them more likely to hear Obama’s public speaking. In a similar justification of a plausible style-shifting context, Soukup (2013) relied on her observation that speakers routinely shift between standard and dialectal Austrian German in stylistic practice. If a speaker evaluation task involves styles that do not coexist in stylistic practice ‘in the wild’ (e.g., the same speaker commanding both an L1 and an L2 accent), the OGT is not likely to work.

Caveats about the OGT notwithstanding, it is clear that traditional approaches to linguistic perception do not give listeners enough credit for being aware of style-shifting; indeed, explicit public awareness of style-shifting (e.g., Meraji, 2013) indicates that listeners may be willing to accept reacting to the same speaker using different features, styles, or languages. Future research should explore the extent to which style-shifting *itself*, not just the individual styles involved in shifting, affects listeners’ judgments of speakers.

6. Conclusion

The current study examined listener ratings of phonetically manipulated speech to test whether listeners were sensitive to such manipulations in the process of making judgments about speaker ethnicity. Regression models indicated that listeners systematically judged a familiar speaker as ‘sounding blacker’ when exposed to more extreme F0 manipulations of both the peak and valley of L + H* pitch accents. This effect was mediated by incrementality, with more extreme L + H* pitch accents mapping to greater perceptions of blackness—albeit with an effect size that suggests caution in generalizing these results. Results of post-hoc testing also reveal that a number of voice quality features appear to also be involved in these judgments. In particular, speech rate, peak delay, HNR, and jitter also appear to influence listener judgments, though the salience of voice quality features may be mediated by the presence versus absence of L + H* pitch accents.

These results have important implications for future work examining both intonational variation from a formal perspective as well as sociophonetic studies on ethnic identification. The finding that listeners seem to attune differently to H* versus L + H* pitch accents in ethnicity judgments and that these perceptions are influenced by phonetic factors provides further motivation for studies that examine intonation from both a phonological and a phonetic perspective. Additionally, the finding that listener perceptions of ethnicity may be manipulated by alterations in F0 provides important context for studies that aim to isolate the phonetic features that may trigger listener judgments of ethnicity. This is especially important given the large body of work on linguistic profiling and discrimination and may provide additional resources for linguists who aim to describe and address racial inequality. Finally, these results indicate that listeners’ sociolinguistic perceptions are sensitive to the magnitude of the input, a finding that indicates promising directions for research in language attitudes and sociolinguistic cognition.

Additional Files

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

- **Appendix A.** Questionnaire. DOI: <https://doi.org/10.5334/labphon.229.s1>
- **Appendix B.** Questionnaire qualitative codes. DOI: <https://doi.org/10.5334/labphon.229.s2>
- **Appendix C.** Voice quality model summaries. DOI: <https://doi.org/10.5334/labphon.229.s3>

Acknowledgements

The authors wish to express their thanks to Paul Reed for comments on the study design. We would also like to thank the audiences at New Ways of Analyzing Variation (NWAV46) and Sociolinguistics Symposium 22, as well as anonymous reviewers for their helpful feedback. Thanks also to our listeners.

Competing Interests

The authors have no competing interests to declare.

References

- Alim, H. S., & Smitherman, G. (2012). *Articulate While Black: Barack Obama, Language, and Race in the U.S.* Oxford: Oxford University Press.
- Beckman, M. E., & Ayers-Elam, G. (1997). Guidelines for ToBI labelling, version 3.0.
- Boersma, P., & Weenink, D. (2015). Praat (Version 5.4.01) [phonetic analysis software]. Available from <http://www.fon.hum.uva.nl/praat/>
- Burdin, R. (2015). Phonological and phonetic variation in list intonation in Jewish English. *Paper presented at NWAV 44*. Toronto. DOI: <https://doi.org/10.21437/SpeechProsody.2014-175>
- Burdin, R., Holliday, N., & Reed, P. (2018). Rising above the standard: Variation in L + H* contour use across 5 varieties of American English. *Paper presented at Speech Prosody*. DOI: <https://doi.org/10.21437/SpeechProsody.2018-72>
- Campbell-Kibler, K. (2009). The nature of sociolinguistic perception. *Language Variation and Change*, 21(1), 135–156. DOI: <https://doi.org/10.1017/S0954394509000052>
- Clopper, C. G. (2010). Phonetic detail, linguistic experience, and the classification of regional language varieties in the United States. In D. R. Preston & N. Niedzielski (Eds.), *A reader in sociophonetics* (pp. 203–222). Berlin: Mouton de Gruyter.
- Clopper, C. G., & Pisoni, D. B. (2007). Free classification of regional dialects of American English. *Journal of Phonetics*, 35(3), 421–438. DOI: <https://doi.org/10.1016/j.wocn.2006.06.001>
- D’Onofrio, A. (2018). Personae and phonetic detail in sociolinguistic signs. *Language in Society*, 47(4), 513–539. DOI: <https://doi.org/10.1017/S0047404518000581>
- Foulkes, P., & Docherty, G. (2006). The social life of phonetics and phonology. *Journal of Phonetics*, 34(4), 409–438. DOI: <https://doi.org/10.1016/j.wocn.2005.08.002>
- Foulkes, P., Docherty, G., Khattab, G., & Yaeger-Dror, M. (2010). Sound judgments: Perception of indexical features in children’s speech. In D. R. Preston & N. Niedzielski (Eds.), *A reader in sociophonetics* (pp. 327–356). Berlin: Mouton de Gruyter.
- Giles, H., & Billings, A. C. (2004). Assessing language attitudes: Speaker evaluation studies. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 187–209). Malden, MA: Blackwell. DOI: <https://doi.org/10.1002/9780470757000.ch7>
- Gumperz, J. J. (1982). *Discourse strategies*. Cambridge: Cambridge University Press. DOI: <https://doi.org/10.1017/CBO9780511611834>
- Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, 48(4), 865–892. DOI: <https://doi.org/10.1515/ling.2010.027>

- Holliday, N. (2016). *Intonational Variation, Linguistic Style, and the Black/Biracial Experience*. (Doctoral dissertation), New York University.
- Holliday, N., & Jagers, Z. S. (2015). Influence of suprasegmental features on perceived ethnicity of American politicians. *Paper presented at 18th International Congress of Phonetic Sciences*, Glasgow, Scotland.
- Jacewicz, E., & Fox, R. A. (2018). The old, the new, and the in-between: Preadolescents' use of stylistic variation in speech in projecting their own identity in a culturally changing environment. *Developmental Science*. DOI: <https://doi.org/10.1111/desc.12722>
- Jun, S.-A., & Foreman, C. (1996). Boundary tones and focus realization in African American English intonations. *Journal of the Acoustical Society of America*, 100(4), 2826. DOI: <https://doi.org/10.1121/1.416648>
- Kaplan, R. (Writer). (2016). Obamas share Super Bowl traditions with Gayle King: CBS News.
- Keating, P., Garellek, M., & Kreiman, J. (2015). Acoustic properties of different kinds of creaky voice. *Paper presented at 18th International Congress of Phonetic Sciences*, Glasgow, Scotland. DOI: <https://doi.org/10.1017/S0025100315000286>
- Kendall, T. (2013). *Speech rate, pause and sociolinguistic variation: Studies in corpus sociophonetics*. New York: Palgrave Macmillan. DOI: <https://doi.org/10.1057/9781137291448>
- Koops, C., Gentry, E., & Pantos, A. (2008). The effect of perceived speaker age on the perception of PIN and PEN vowels in Houston, Texas. *University of Pennsylvania Working Papers in Linguistics*, 14(2).
- Kuznetsova, A., Brockhoff, B., & Christensen, H. B. (2016). lmerTest (Version 2.0-33) [R package]. Available from <https://CRAN.R-project.org/package=lmerTest>
- 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., Ash, S., Ravindranath, M., Weldon, T., Baranowski, M., & Nagy, N. (2011). Properties of the sociolinguistic monitor. *Journal of Sociolinguistics*, 15(4), 431–463. DOI: <https://doi.org/10.1111/j.1467-9841.2011.00504.x>
- Leach, H., Watson, K., & Gnevsheva, K. (2016). Perceptual dialectology in northern England: Accent recognition, geographical proximity and cultural prominence. *Journal of Sociolinguistics*, 20(2), 192–211. DOI: <https://doi.org/10.1111/josl.12178>
- Lefcheck, J. S. (2016). piecewiseSEM: Piecewise structural equation modelling in R for ecology, evolution, and systematics. *Methods in Ecology and Evolution*, 7(5), 573–579. DOI: <https://doi.org/10.1111/2041-210X.12512>
- Loman, B. (1975). Prosodic patterns in a Negro American dialect. In H. Ringbom, A. Ingberg, R. Norrman, K. Nyholm, R. Westman & K. Wikberg (Eds.), *Style and text: Studies presented to Nils Erik Enkvist* (pp. 219–242). Stockholm: Språkförlaget Skriptor AB.
- McLarty, J. (2018). African American Language and European American English intonation variation over time in the American South. *American Speech*, 93(1), 32–78. DOI: <https://doi.org/10.1215/00031283-6904032>
- McLarty, J., Vaughn, C., & Kendall, T. (2017). Acoustic correlates of perceived prosodic prominence in African American English and European American English. *Paper presented at NWAV 46*, Madison, Wisconsin.
- Meraji, S. M. (2013). Why Chaucer said 'ax' instead of 'ask,' and why some still do. *National Public Radio*. Retrieved from <https://www.npr.org/sections/codeswitch/2013/12/03/248515217/why-chaucer-said-ax-instead-of-ask-and-why-some-still-do>

- Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology, 18*(1), 62–85. DOI: <https://doi.org/10.1177/0261927X99018001005>
- Pharao, N., Maegaard, M., Møller, J. S., & Kristiansen, T. (2014). Indexical meanings of [s+] among Copenhagen youth: Social perception of a phonetic variant in different prosodic contexts. *Language in Society, 43*(1), 1–31. DOI: <https://doi.org/10.1017/S0047404513000857>
- Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation*. (Doctoral dissertation), Massachusetts Institute of Technology, Cambridge, MA.
- Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics, 2*(1). DOI: <https://doi.org/10.1146/annurev-linguistics-030514-125050>
- Pierrehumbert, J. B., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan & M. Pollack (Eds.), *Intentions in communication* (pp. 271–311). Cambridge, MA: MIT Press.
- Plichta, B., & Preston, D. R. (2005). The /ay/s have it: The perception of /ay/ as a North-South stereotype in US English. *Acta Linguistica Hafniensia, 37*, 243–285. DOI: <https://doi.org/10.1080/03740463.2005.10416086>
- Podesva, R. J. (2011). Salience and the social meaning of declarative contours. *Journal of English Linguistics, 39*(3), 233–264. DOI: <https://doi.org/10.1177/0075424211405161>
- Purnell, T., Idsardi, W. J., & Baugh, J. (1999). Perceptual and phonetic experiments on American English dialect identification. *Journal of Language and Social Psychology, 18*(1), 10–30. DOI: <https://doi.org/10.1177/0261927X99018001002>
- R Core Team. (2018). R: A language and environment for statistical computing (Version 3.5.2). Vienna. Available from <https://www.R-project.org/>
- Reed, P. (2016). *Sounding Appalachian: /aI/ Monophthongization, Rising Pitch Accents, and Rootedness*. (Doctoral dissertation), University of South Carolina.
- Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin, 2*(6), 110–114. DOI: <https://doi.org/10.2307/3002019>
- Soukup, B. (2013). ‘Matched guise technique’ vs. ‘Open guise technique’ in the elicitation of language attitudes: Insights from a comparative study. *Paper presented at ExAPP 2, Copenhagen*.
- Szakay, A. (2012). Voice quality as a marker of ethnicity in New Zealand: From acoustics to perception. *Journal of Sociolinguistics, 16*(3), 382–397. DOI: <https://doi.org/10.1111/j.1467-9841.2012.00537.x>
- Tarone, E. (1973). Aspects of intonation in Black English. *American Speech, 48*(1/2), 29–36. DOI: <https://doi.org/10.2307/3087890>
- Thomas, E. R. (2011). *Sociophonetics: An introduction*. New York: Palgrave Macmillan. DOI: <https://doi.org/10.1007/978-1-137-28561-4>
- Thomas, E. R. (2015). Prosodic features of African American English. In S. Lanehart (Ed.), *The Oxford handbook of African American Language* (pp. 420–438). Oxford: Oxford University Press.
- Thomas, E. R., & Reaser, J. (2004). Delimiting perceptual cues used for the ethnic labeling of African American and European American voices. *Journal of Sociolinguistics, 8*(1), 54–87. DOI: <https://doi.org/10.1111/j.1467-9841.2004.00251.x>
- Todd, R. (2002). Speaker-ethnicity: Attributions based on the use of prosodic cues. *Paper presented at Speech Prosody, Aix-en-Provence, France*.
- Tyler, J. C. (2015). Expanding and mapping the indexical field: Rising pitch, the uptalk stereotype, and perceptual variation. *Journal of English Linguistics, 43*(4), 284–310. DOI: <https://doi.org/10.1177/0075424215607061>

- Villarreal, D. (2016). "Do I sound like a Valley Girl to you?" Perceptual dialectology and language attitudes in California. In V. Fridland, T. Kendall, B. Evans & A. Wassink (Eds.), *Speech in the Western states* (Vol. 1, pp. 55–75). Durham, NC: Duke University Press. DOI: <https://doi.org/10.1215/00031283-3772901>
- Villarreal, D. (2018). The construction of social meaning: A matched-guise investigation of the California Vowel Shift. *Journal of English Linguistics*, 46(1), 52–78. DOI: <https://doi.org/10.1177/0075424217753520>
- Zahn, C. J., & Hopper, R. (1985). Measuring language attitudes: The speech evaluation instrument. *Journal of Language and Social Psychology*, 4(2), 113–123. DOI: <https://doi.org/10.1177/0261927X8500400203>

How to cite this article: Holliday, N., & Villarreal, D. 2020 Intonational variation and incrementality in listener judgments of ethnicity. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 11(1):3, pp. 1–21. DOI: <https://doi.org/10.5334/labphon.229>

Submitted: 22 September 2019 **Accepted:** 20 February 2020 **Published:** 01 April 2020

Copyright: © 2020 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

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

OPEN ACCESS 