An investigation of the systematic use of spectral information in the determination of apparent-talker height

Santiago Barreda[a)]
Department of Linguistics, University of California, Davis, Davis, California 95616, USA

(Received 19 August 2016; revised 22 May 2017; accepted 25 May 2017; published online 28 June 2017)

The perception of apparent-talker height is mostly determined by the fundamental frequency (f0) and spectral characteristics of a voice. Although it is traditionally thought that spectral cues affect apparent-talker height by influencing apparent vocal-tract length, a recent experiment [Barreda (2016). J. Phon. **55**, 1–18] suggests that apparent-talker height can vary significantly within-talker on the basis of phonemically-determined spectral variability. In this experiment, listeners were asked to estimate the height of 10 female talkers based on manipulated natural productions of bVd words containing one of /i æ ɑ u ɝ/. Results indicate that although listeners appear to use vocal-tract length estimates in determining apparent-height, apparent-talker height also varies significantly within-talker based on the inherent spectral and source characteristics of different vowels, with vowels with lower formant-frequencies and f0 being associated with taller talkers overall. The use of spectral and f0 information in apparent-height estimation varied considerably between listeners, resulting in additional variation in the apparent-height of talkers. Although the use of acoustic information in the determination of apparent-height was highly systematic, it does not necessarily follow from the empirical relationship between speech acoustics and actual talker height.
 [http://dx.doi.org/10.1121/1.4985192]

[TCB] Pages: 4781–4792

I. INTRODUCTION

Taller talkers tend to produce speech with a lower average fundamental frequency (f0) and formant frequencies than shorter talkers, across the entire human population. Listeners show a sensitivity to this covariation and consistently associate voices with lower f0s and formant frequencies with taller talkers (Collins, 2000; Rendall *et al*., 2007; van Dommelen and Moxness, 1995). Although this strategy is useful for determining the relative heights of talkers across different general talker classes (e.g., adults vs children, men vs women), it has long been noted that judgments of talker height from speech are not particularly accurate if one controls for age and sex.[1] However, despite this inaccuracy, listeners are systematic and predictable in their judgments of talker height (Collins, 2000; Rendall *et al*., 2007; van Dommelen and Moxness, 1995).

The use of f0 information in height estimation appears to be relatively straightforward: listeners consistently associate lower f0s with taller talkers. However, because of its crucial role in signaling phonemic contrasts, the use of spectral information (e.g., formant frequencies) in the determination of apparent-talker height may be substantially more complicated. The focus of this work will be on the systematic use of spectral information by human listeners when estimating apparent-talker height.

A. The systematic use of spectral information in height estimation

1. Vocal-tract length based height estimation

Taller talkers tend to have longer vocal tracts (Fitch and Giedd, 1999) and, as a result, produce lower formant frequencies overall. As a result, low formant frequencies, and a longer apparent vocal-tract length (VTL), can be evidence of a taller talker.

Vocal-tract related differences between-talkers of a dialect manifest primarily as uniform multiplicative increases or decreases (i.e., uniform scaling) of the formant-patterns produced for a given vowel category (Barreda, 2016; Barreda and Nearey, 2013; Nearey, 1978; Turner *et al*., 2009). Uniform scaling of formant patterns means that if two productions of a vowel phoneme produced by two talkers differ by 10% in their F1 frequencies, they are expected to differ by roughly the same amount in their F2 and F3 frequencies, on average.

Spectral information in speech sounds is usually thought to affect apparent-talker height by informing an estimate of the VTL[2] of the talker. For example, van Dommelen and Moxness (1995, p. 283) state that "large [vocal-tract] values, that is low formant frequencies, were interpreted by the listeners as indicating large body dimensions," and Rendall *et al*. (2007, p. 1215) suggest that listeners "discriminate size differences based on formant frequency cues to speaker VTL." From this perspective, formant frequencies inform apparent-talker height only by affecting the listener's estimate of the VTL of the talker (i.e., apparent VTL). However, VTL information is conflated with phonemically determined spectral variability in the speech signal.

For example, consider the tokens of /o/ produced by talkers with long (circle) and short (cross) VTLs in Fig. 1. When phonetic content is controlled for, lower formant frequencies can be taken as direct evidence of a longer VTL. However, advancement in the direction typically denoting increases in VTL cannot always be interpreted as direct evidence of a long VTL. Talkers that produce vowels near the circle in the /ɑ/ distribution will have long vocal-tracts, and

a)Electronic mail: sbarreda@ucdavis.edu

yet they are very close to short-VTL tokens of /o/ (cross). Similar situations arise when one compares the absolute locations of the example tokens on Fig. 1 without controlling for phonetic category.

2. Theories supporting the availability of phoneme-independent estimates of vocal-tract length

In order to arrive at VTL-based height judgments from a small set of speech sounds, listeners would need to separate phoneme-dependent, and speaker-dependent (i.e., VTL) information in the absolute formant-pattern. There are three general theories of speech perception that suggest that phoneme-independent VTL estimates (or analogous information) may be available to listeners as a by-product of speech perception. Although the theories to be outlined differ substantially in their mechanisms, all three suggest that listeners should be able to make apparent-talker height judgments that vary systematically with respect to talker VTL, while ignoring spectral information related to phoneme identity.

The first class of theories suggest that listeners identify vowel sounds on the basis of the relative position of the vowel sound within the vowel space of the apparent talker (Joos, 1948; Ladefoged and Broadbent, 1957; Nearey, 1978). Given that the vowel spaces of talkers of a dialect differ primarily according to VTL, having expectations regarding the talker's vowel space entails having something like a VTL estimate. From this perspective, in order to identify an ambiguous point between the long-VTL /ɑ/ and the short-VTL /o/ in Fig. 1, the listener must decide if the talker has a long or short VTL, which will then determine the interpretation of the vowel. As a result, the relative position of a token in a formant space with respect to other tokens of the same phoneme can be informative as to the VTL of the talker who

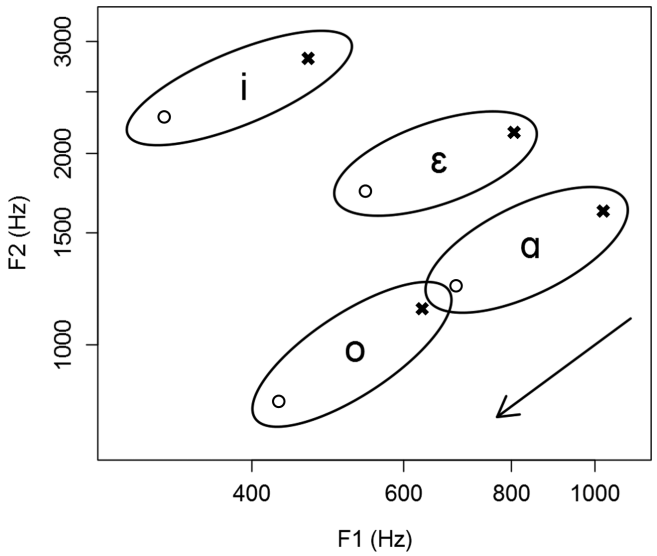

FIG. 1. Distribution of selected vowels from Hillenbrand *et al.* (1995). Ellipses enclose two standard deviations. The arrow indicates the direction of variation according to uniform, multiplicative increases in formant patterns of the kind associated with vocal-tract length increases. In general, talkers with longer vocal-tracts will produce tokens near the circles, while talkers with shorter vocal-tracts will produce tokens towards the crosses.

produced them. Given that this class of theories do not directly seek to explain height estimation, they do not make strong predictions about this behavior. However, if listeners do identify vowels in a talker-dependent manner and use resulting VTL estimates when estimating talker height, then we should expect an effect for talker VTL on a trial-by trial basis.

The second group of theories suggest that listeners identify speech sounds on the basis of previously heard exemplars stored in memory (Goldinger, 1998; Johnson *et al.*, 1999). These theories suggest that knowledge of the acoustic characteristics of phonemes is intrinsically associated with knowledge of the indexical characteristics (e.g., sex, age, height) of the speakers that produce them. This indexical information would include the average spectral scaling (i.e., VTL) associated with the formant patterns produced by that speaker, at least implicitly, by virtue of recognizing that a given talker produces high/low formant frequencies overall. As a result, although proponents of these theories may not consider that listeners estimate speaker VTL in speech perception, they would suggest that listeners have access to something like a phoneme-independent spectral scaling estimate analogous to VTL for the purposes of apparent-height estimation.

Exemplar-based explanations of talker-height perception rely on associations between experienced exemplars and the indexical characteristics of the talkers that produced them. Talkers do not vary by height according to the phonemes they produce, and so listeners should not have made any associations between specific phonemes and different apparent-heights, within talker. Consequently, there should not be any systematic within-talker variability in height judgments across different the phonemes produced by a talker.

Finally, it has been suggested that "the auditory system includes an active re-scaling process that is applied to all sounds at an early point in the auditory system […] thereby reducing variability in the [spectral] shape information whilst segregating size information" (Turner *et al.*, 2006, p. 154). According to these theories, phoneme-independent VTL estimates are easily available to human listeners and are hypothesized to form the basis of height estimation by human listeners (Irino and Patterson, 2002; Ives *et al.*, 2005; Smith and Patterson, 2005; Turner *et al.*, 2006). In light of these strong claims about the separation of VTL and phoneme-dependent information at an early stage in auditory processing, any systematic variability in apparent height across phonemes poses a significant challenge to this group of theories.

3. Phoneme biases and within-talker variation in apparent height

Researchers investigating the use of spectral information in height estimation typically control for the phonetic content of the speech sounds being considered, either by presenting listeners with pairs of phonetically identical stimuli, or by only considering results aggregated over sets of stimuli with the same phonetic content (Ives *et al.*, 2005; Rendall

et al., 2007; Smith and Patterson, 2005). As noted in Barreda (2016), focusing on same-phoneme comparisons will give the impression that height estimation is VTL-based even if listeners also respond to phoneme-specific spectral information in height estimation. For example, if listeners were only asked to compare the vowels presented in Fig. 1 within-phoneme, either responding to VTL cues or simply responding to the absolute spectral characteristics of the sounds would both result in the long-VTL vowels being associated with taller talkers. On the other hand, if listeners were asked to compare phonemes from different categories (e.g., long-VTL /ɑ/ vs short-VTL /o/) then listeners would really need to correct for phonetic-content in order to identify the long-VTL talker as taller.

Barreda (2016) presented listeners with pairs of synthetic vowels from the set of /i æ ʊ/, and asked listeners to identify the taller talker. Vowels varied on the basis of vowel quality and/or simulated VTL differences, but were matched for all other acoustic characteristics. Listeners demonstrated a tendency to identify /ʊ/ as being produced by a taller apparent-talker than /æ/, independently of the VTL implied by the vowels. Barreda (2016) suggested that listeners associate /ʊ/ with taller talkers due to the substantially lower F1 and F2 of /ʊ/ relative to /æ/. The associations between specific phonemes and taller or shorter apparent talkers were termed "phoneme biases" in height perception. We might imagine the "true" apparent height of a talker estimated on the basis of all the speech sounds produced by the talker. The term phoneme bias is used here to represent systematic differences between apparent height based on a single or limited set of phonemes, and the "true" apparent height of a talker.

B. The current experiment

Apparent-height judgments for a set of talkers are usually investigated using estimates of talker VTL (Collins, 2000; Fitch, 1994; Ives *et al.*, 2005; Rendall *et al.*, 2007; Smith and Patterson, 2005; van Dommelen and Moxness, 1995). However, the presence of phoneme biases in height perception suggests a more complicated use of spectral information in the estimation of apparent-talker height. Although it seems that apparent-talker height can vary systematically within-talker, it is not clear how large phoneme-bias effects are for natural voices that vary along many acoustic dimensions simultaneously, or for longer stretches of speech than isolated vowels, where more information about talker height is present.

The goal of the experiment is to investigate phoneme biases in apparent-height in a less controlled setting by using several real voices, and asking listeners to make absolute height-judgments. Listeners heard bVd words containing five different vowels produced by ten adult female talkers and were asked to estimate the absolute height of the talkers in feet and inches. In addition to featuring natural variation in VTL between different talkers, the experiment also included simulated VTL differences in order to increase the amount of spectral variability between talkers, and to compare the effects of natural and simulated variability in VTL. Spectral sources of variance are of primary interest, both between (real and simulated VTL), and within-talker (phoneme biases). Importantly, these effects will be estimated independently of within- and between-talker variation in f0.

II. METHOD

A. Participants

Participants were 38 undergraduate students (12 males, 26 females) from the University of California, Davis. All listeners were native English-speakers and reported no known hearing problems. Listeners participated in exchange for partial course credit.

B. Stimulus information

Stimuli were bVd words containing the vowels /æ ɑ ɝ u i/ ("bad," "bod," "bird," "booed," "bead"), produced by ten adult female native-speakers of California English, ranging in height from 61 to 69 in. [mean = 65.75, standard deviation (sd) = 2.64]. Stimulus words were recorded in a sound-attenuated booth, produced in isolation and in a random order. Only female talkers were used in order to avoid gender judgments from complicating the relationship between apparent height and speech acoustics, especially in light of the acoustic similarities between adult females and prepubescent male talkers.

Figure 2 presents information about the stimulus words used in the experiment, and average formant frequencies for each vowel are presented in Table I. As seen in Fig. 2(b), there appears to be an effect for intrinsic vowel f0, with vowels with a higher F1 having lower f0s on average (Whalen and Levitt, 1995). The geometric mean of the first three formant frequencies (GMFF) produced by each talker across their five representative vowels was calculated in order to estimate differences in VTL between the talkers (Nearey, 1978). Figure 2(c) presents GMFF and mean log-f0 for each talker. The voice with the highest GMFF had formant-frequencies that were 12% higher on average than those of the voice with the lowest GMFF (1331 and 1491 Hz), while average f0s spanned from 184 to 251 Hz, a difference of 36%. These ranges are reasonable given the amount of variability seen in large datasets. For example, for the 48 adult-female talkers in Hillenbrand *et al.* (1995), the highest GMFF is 20% higher than the lowest, while the highest average f0 is 59% higher than the lowest.

Simulated VTL differences were carried out using the "change gender" function in Praat (Boersma and Weenink, 2001). All 50 stimulus words (5 words for each of 10 talkers) were scaled up by a factor of 1.06 and down by 1/1.06, resulting in a scaling difference of 12.36% between the two simulated VTL levels (long and short). The 6% changes in VTL are close to just noticeable differences in two-alternative forced choice tasks, estimated to be between 1% and 6% (Charlton *et al.*, 2013; Ives *et al.*, 2005; Pisanski and Rendall, 2011).

C. Procedure

Sounds were presented over Sennheiser HD 280 headphones, in a sound-attenuated booth. For each trial, listeners

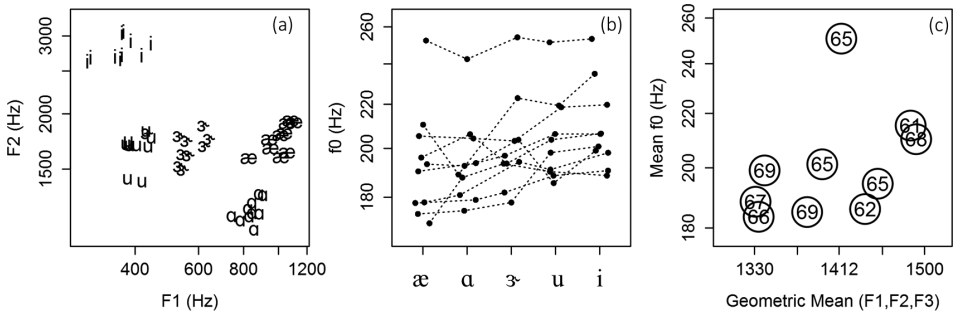

FIG. 2. (a) First two formant frequencies at vowel midpoints for each stimulus word. (b) Distribution of mean f0 by vowel phoneme, for all stimuli. (c) The ten stimulus talkers presented according to geometric-mean formant frequency (GMFF) for the first three formants for all vowels, and mean log-f0 across all vowels. Numbers indicate talker height in inches.

were presented with a single word and asked to indicate the height of the talker by clicking on a ruler presented on a computer monitor. The ruler ranged from 4 ft 6 in. to 6 ft 0 in., which includes 99.84% of adult females (mean = 63.8 in., sd = 2.74 in., Fryar *et al.*, 2012). After each click on the ruler, text above the ruler indicated the height associated with the click, rounded to the nearest tenth of an inch. Listeners could replay the sound up to two additional times per trial. Once listeners were satisfied with their response, they were asked to press a button marked "Submit," and the following stimulus word played after a 1 s pause. Responses could take on any integer value between 1 and 720, according to the pixel associated with the final response provided by the listener. Participation was limited to 30 min. Listeners were presented with each of the 100 stimuli twice, blocked by repetition but randomized along all other stimulus dimensions. All listeners provided 200 responses save for one listener who only completed 132 trials, for a total of 7532 responses.

D. Statistical analysis: Bayesian multilevel linear-regression

Results were analyzed using a Bayesian multilevel linear-regression model. Individual height responses were analyzed as coming from a normal distribution with an unknown mean and listener-specific error-term (for listener l) as in Eq. (1). The mean parameter (μ) was broken down in an analysis of variance (ANOVA)–style decomposition consisting of the following intercept terms [Eq. (2)]: an overall intercept (α_0), Talker (T, 10 levels), Listener (L, 38 levels), Vowel (V, 5 levels), the Listener ×Vowel interaction (LV, 190 levels), and the Listener ×Talker interaction (LT, 380 levels)

$$y \sim N(\mu, \sigma_l^2), \tag{1}$$

$$\mu = \alpha_0 + T + L + V + LV + LT + \beta_{cf0}cf0 + \beta_{VTL}VTL. \tag{2}$$

TABLE I. Average formant frequencies at vowel midpoints for each of the stimulus vowel phonemes across all talkers.

	æ	i	ɑ	ɜ˞	u
F1	1008	366	839	556	403
F2	1754	2805	1204	1686	1660
F3	2843	3289	2887	1977	2768

Height responses also varied as a linear function of the simulated vocal-tract length level of the voice (VTL), and vowel f0 in log-Hz, centered within-talker (cf0). Vowel f0 was centered so that it would reflect within-talker variation in f0 between different vowel tokens, rather than between-talker variation in average f0. The slope terms were decomposed into intercept (β^0_{cf0}, β^0_{VTL}) and listener-deflection terms (β^l_{cf0}, β^l_{VTL}) as in Eqs. (3) and (4). The listener deflection terms are equivalent to random slopes for each listener, while the intercept terms are equivalent to the fixed effects for the predictors

$$\beta_{cf0} = \beta^0_{cf0} + \beta^l_{cf0}, \tag{3}$$

$$\beta_{VTL} = \beta^0_{VTL} + \beta^l_{VTL}. \tag{4}$$

The ten Talker coefficients in Eq. (2) represent average height judgments for each talker across all listeners. The average apparent-height for each talker was modeled on the basis of their GMFF in log-Hz and mean log-f0 (mf0), as in Eq. (5). The 380 Listener × Talker interaction coefficients in Eq. (2) represent the listener-specific height judgments for Listener l and Talker t. These coefficients were also modeled as varying on the basis of talker GMFF and log-mean f0, however, in this case the coefficients for each predictor were allowed to vary between listeners as in Eq. (6)

$$T_t = \beta_{mf0}mf0_t + \beta_{GMFF}GMFF_t, \tag{5}$$

$$LT_{lt} = \beta^l_{mf0}mf0_t + \beta^l_{GMFF}GMFF_t. \tag{6}$$

Each coefficient with a single degree of freedom was given a normal prior with a mean of 0 and a variance of 100. Each group of coefficients with more than one degree of freedom, was modeled as coming from a higher-level normal distribution with a mean of zero, and variance parameters that were estimated from the data. All higher-level population variance parameters were given half-Cauchy priors with location and scale parameters of 0 and 5, respectively.

Posterior samples for all parameters were generated using JAGS (Plummer, 2003) and R (R Core Team, 2015). Credible intervals for parameters, or combinations of parameters, will be assessed using the 95% highest-density interval (HDI; Kruschke, 2010), representing the interval enclosing 95% of the posterior distribution such that every value inside the interval is more probable than every value outside the interval. Posterior distributions will also be characterized using their mean values, and the

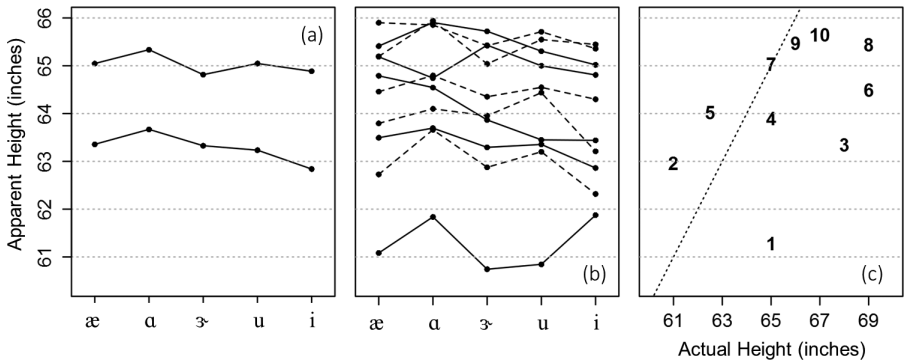

FIG. 3. (a) Average reported height for each vowel, for short (lower line) and long (upper line) VTL levels. (b) Average reported height by talker and vowel, averaged across all listeners. Line types alternate only to improve legibility. (c) Average apparent-height plotted against the actual height of each of the ten talkers who provided stimuli. Talker numbers indicate apparent-height rank, with lower numbers indicating shorter apparent talkers.

percentage of the distribution above or below zero where appropriate.

Responses were standardized within-listener prior to analysis in order to remove between-listener variation in the use of the response scale. Standardized height values are useful for directly comparing the use of acoustic information by different listeners, but do not result in a scale that is easy to interpret. To remedy this, results will be presented in inches based on the average mean and standard deviation across all listeners. Reported effects may be converted back to standardized height judgments by dividing the value of interest by 2.47, the standard deviation of height judgments across all listeners in the experiment.

In the model outlined above, the slope coefficients for mean log-f0 and GMFF reflect the effect of a one unit log-Hz change in log-f0 or GMFF on apparent-talker height. However, a change of this magnitude is larger than the adult female range of mean log-f0, and considerably larger than the range of GMFF across all humans. For this reason, GMFF and f0 effects were multiplied by 0.1165 [ln(1.1236)] so that they would reflect the effect of a 12.36% change in these cues on apparent height, the magnitude of the spectral shifts used to simulate VTL differences in this experiment. As a result, the reported effects reflect reasonable ranges of variation for adult females, and the effects of f0, GMFF and simulated VTL are directly comparable since they represent changes of equal magnitude.

III. RESULTS

Figure 3(a) presents average apparent-height for each vowel at each VTL scaling level, showing effects for VTL differences, and a clear pattern of phoneme biases at both scaling levels. Figure 3(b) indicates that there was a considerable amount of between-talker variability in apparent height, and some talkers were consistently identified as taller than others overall. However, there is also a considerable amount of within-talker variability in apparent height on the basis of vowel category such that the perceived relative height of pairs of talkers could depend on the vowels being considered. As seen in Fig. 3(c), actual height was not very predictive of average apparent-talker height for this sample of talkers, with 17% shared variance ($r = 0.41$) between the two values.[3]

Results will be analyzed using the model described in Sec. II D. The relative contribution of different predictors to variation in apparent-talker height will be inspected using a Bayesian Analysis of Variance (Gelman, 2005), presented in Fig. 4. The Bayesian ANOVA approach estimates the standard deviation of each batch of predictors (e.g., Talker, Vowel) using the posterior distribution of these parameter estimates. If one group of predictors has a larger standard deviation than a second group, then it is a larger contributor, on average, to variation in the dependent variable. For example, between-talker variability in Fig. 3(b) is much larger than between-vowel variability in Fig. 3(a), and this discrepancy is reflected by the difference in the estimated standard deviations of the Talker and Vowel effects in Fig. 4(a).

The largest effect was for simulated VTL differences (a spectral scale difference of 12.4%) resulting in a perceived height difference of −1.73 in. (95% HDI = [−1.813, −1.655]). The large effect for VTL indicates that the simulated VTL shifts worked as intended, with lower spectral

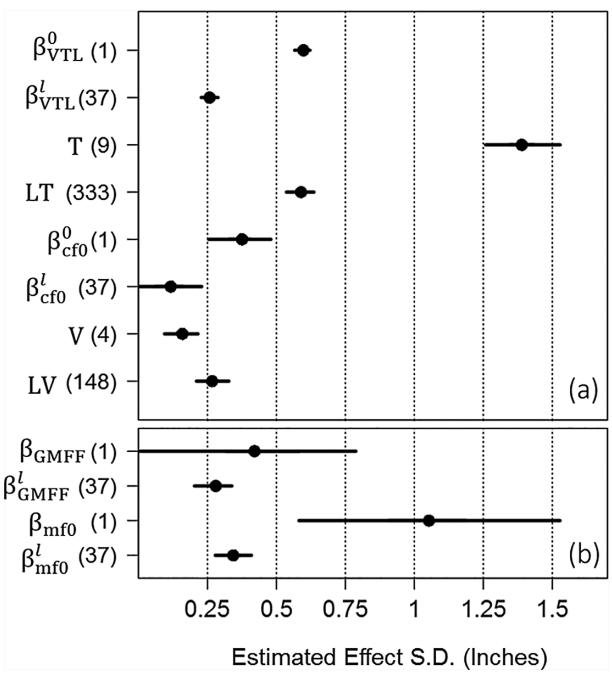

FIG. 4. Posterior distribution of standard deviation estimates for the predictors presented in Eqs. (2)–(6) for (a) data-level and (b) talker-level predictors. Points indicate means of posterior distributions and lines indicate the 95% HDI for each standard deviation estimate. Numbers in brackets indicate the degrees of freedom for each estimate. For predictors with a single degree of freedom, the posterior distribution of absolute values is shown.

scaling (and an implied longer VTL) being associated with taller talkers overall.

A. Between-talker variability in apparent height

The second-largest source of variance in apparent-talker height is due to differences between talkers. Average Talker effects were highly predictable on the basis of the mean log-f0 and GMFF of each talker,[4] with an estimated mean R^2 of 0.84 (95% HDI = [0.707, 0.928]). Effects for mean log-f0 (β_{mf0}) and GMFF (β_{GMFF}) were roughly equal in magnitude, though GMFF (Mean = −1.238; 95% HDI = [−2.536, 0.073]; 96.9% < 0) had a much broader credible interval than mean log-f0 (Mean = −1.293; 95% HDI = [−1.896, −0.722]; 99.9% < 0). The uncertainty in the estimation of the GMFF and mean log-f0 parameters (seen in Fig. 4) is a reflection of the fact that they were estimated using only ten unique talkers.

B. Within-talker variability in perceived height

1. Within-talker variability based on f0

The effect of within-talker variability in f0 on apparent-talker height was estimated using centered log-f0, which indicated how much above or below the talker's mean log-f0 a given production was. A centered log-f0 difference of 12.4% resulted in a perceived height change of −0.459 in. (95% HDI = [−0.595, −0.316]; 100% < 0), a value about one third as large as that of mean log-f0. The effect for centered log-f0 indicates that lower f0s are associated with taller heights even across repeated utterances by a single talker.

2. Vowel-specific spectral pattern

Figure 5 presents average apparent-height judgments for each phoneme, averaged across all listeners and talkers (dashed line). These between-phoneme differences in apparent height represent the sum of intrinsic f0 and spectral characteristics for each vowel phoneme. The independent effect of phoneme-specific spectral information on apparent-height may be considered using the model vowel-category coefficients [V, Eq. (5)], since these coefficients are estimated independently of centered log-f0. The Vowel terms presented in Fig. 5 show substantial variation between phonemes, independently of the intrinsic f0 patterns between vowels. The largest difference between vowels based only on spectral information was a 0.33-in. difference between /æ/ and /ɑ/ (95% HDI = [−0.497, −0.164]).

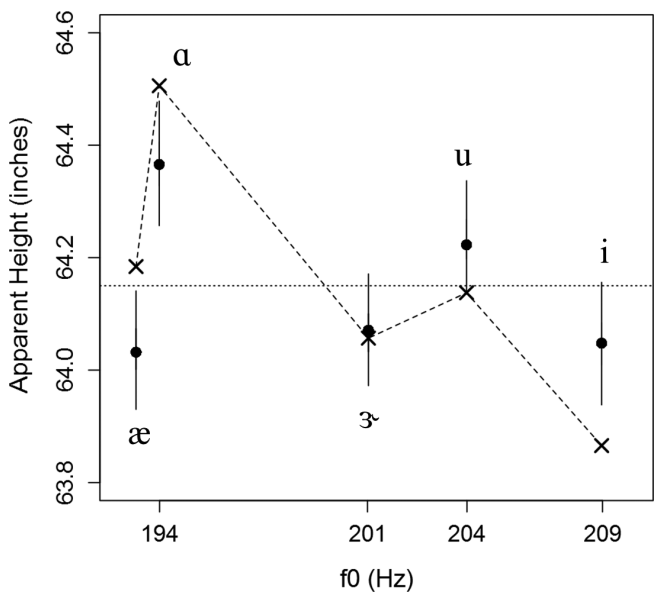

FIG. 5. Average height reported for each vowel across all listeners and talkers (dashed line). Filled points indicate mean vowel effects, centered around the mean reported height. Vertical lines indicate 95% HDIs for each vowel effect.

The Listener × Vowel interaction [LV, Eq. (2)] captures variability in spectrally based phoneme biases across different listeners. It is the only main-effect term in Fig. 4(a) for which the listener-by-effect interaction is of a larger magnitude than the main effect itself. However, it seems that the smaller vowel main effect is partly attributable to the fact that three of the vowels considered (/æ ɜ˞ i/) are associated with talkers of roughly equal height. If the experiment had only featured /i/ and /ɑ/, for example, the mean effect for vowel category would have been larger than the between subject variability. A visual inspection of the listener-specific vowel effects indicated that between-listener variability does not overwhelm or obscure the pattern of effects seen in Fig. 5.

C. Between-listener variation in talker-height estimation

As seen in Fig. 4, there was substantial between-listener variability in sensitivity to simulated VTL (β^{l}_{VTL}), GMFF ($\beta^{l}_{\mathrm{GMFF}}$) and mean log-f0 ($\beta^{l}_{\mathrm{mf0}}$). However, there was very little variability in the use of centered log-f0 (β^{l}_{cf0}) across listeners, and a large amount of the posterior density of the standard deviation was concentrated near zero. As a result, the discussion to follow will focus on between-listener variability in GMFF, log-mean f0, and simulated VTL.

Figure 6 presents listener-specific effects for simulated VTL, GMFF, and mean log-f0. GMFF and VTL effects are positively correlated across listeners (Mean = 0.673; 95% HDI = [0.531, 0.808]), indicating that listener sensitivity to GMFF is predictive of sensitivity to simulated VTL shifts. Thus, it appears that much of the between-listener variability in Fig. 6(a) may reflect differences in the ability to estimate, use, or report apparent-VTL information, rather than simply being noise.

1. Modeling the talker by listener interaction

The Listener × Talker interaction term in Eq. (2) allows for the listener-specific talker judgments to be modeled, rather than simply finding the average apparent-height for each talker across all listeners. As seen by the magnitude of the Listener × Talker effect in Fig. 4(a) (LT), between-listener variability in height judgments for individual talkers represents a substantial component of the variation in apparent-talker height.

The Listener × Talker interaction terms were very predictable on the basis of listener-specific usage of talker mean log-f0 and GMFF, with 52% of the variance being accounted

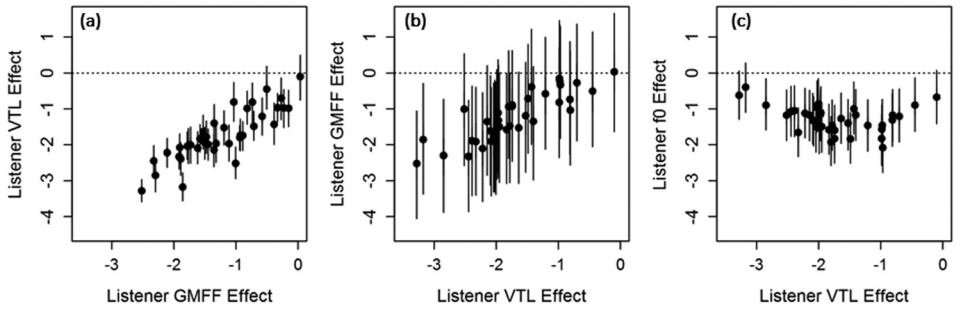

FIG. 6. Estimates of listener-specific effects are presented for (a) simulated VTL plotted against the effect for GMFF, (b) GMFF plotted against simulated VTL (c) mean log-f0 plotted against simulated VTL. Points indicate means, vertical lines indicate the 95% HDI for parameter estimates.

for (95% HDI = [40.9%, 62.8%]). The large amount of variance explained indicates that much of the variation in the apparent height of individual talkers for different listeners arises from systematic variability in the use of f0 and GMFF between listeners, rather than representing unconstrained, idiosyncratic differences in the preferences of different listeners.

IV. DISCUSSION

A. Estimation of apparent vocal-tract length in height perception

Most listeners consistently associate tokens at the long-VTL level with taller talkers independently of vowel and f0 information. Although the credible interval of the effect for β_{GMFF} was rather wide, the mean posterior estimate was similar to that of β_{VTL}, and listener-specific estimates for GMFF and VTL effects (β^{l}_{GMFF}, β^{l}_{VTL}) were strongly correlated. The similarity of the effects for natural and simulated VTL differences is a reassuring finding given that linearly scaling the spectral envelope of speech sounds to simulate VTL differences between talkers is a commonly-used method, and underlies much of the research into human perception of apparent-height (Barreda, 2016; Charlton *et al.*, 2013; Ives *et al.*, 2005; Pisanski and Rendall, 2011; Rendall *et al.*, 2007; Smith and Patterson, 2005).

If listeners were associating the long-VTL level with taller talkers only because of lower absolute formant-frequencies, then the effect for GMFF should be much smaller than the effect for Vowel, since the 12.4% changes in formant frequencies between the VTL levels are much smaller than the between-phoneme variation in formant patterns. Instead, because of the relative sizes of the VTL and spectral vowel-effects, and the amount of between-talker variation in VTL, the apparent height of a talker across different phonemes will tend to cluster around a value determined by the apparent-VTL of the talker, as shown in Fig. 7. Such a result is consistent with vocal-tract length based height estimation.

1. Evaluation of support for theories supporting vocal-tract length estimation

Although listeners appear to use information about the average spectral scaling associated with a given talker (i.e., VTL) when assessing talker height from a single word, this information is not directly available in the absolute formant-pattern present in any given speech sound. The three classes of theories of speech perception that support VTL-estimation (Sec. I A 2) will be discussed here in terms of the results.

According to the first group of theories, listeners can use the relative locations of vowel tokens with respect to other vowels of the same category in order to estimate talker VTL (Nearey, 1978; Ladefoged and Broadbent, 1957). For example, we may consider the GMFF of each stimulus (for F1, F2 and F3), relative to the average GMFF of stimuli of the same category, effectively a pattern-corrected GMFF. The pattern-corrected GMFF indicates whether the formant-pattern for a stimulus is high or low independently of phonetic category, and is therefore related to talker VTL. Although VTL information was not directly available in any given trial, there was a correlation of 0.81 between the overall GMFF of each talker in this experiment and the pattern-corrected GMFF for each stimulus produced by a talker. As a result, it seems plausible that the listeners in this experiment could be inferring the apparent-VTL of talkers in the manner suggested by the first group of theories.

Since the first class of theories were intended to explain listener adaptation to apparent-talker VTL for the purposes of vowel perception, they do not make strong claims about the use of apparent-VTL estimates in determining apparent-height, nor about the use of other spectral information in apparent-height estimates. As such, the first class of theories is generally compatible with all of the results presented here, although this is in part due to making fewer and weaker predictions than the other theories to be discussed. However,

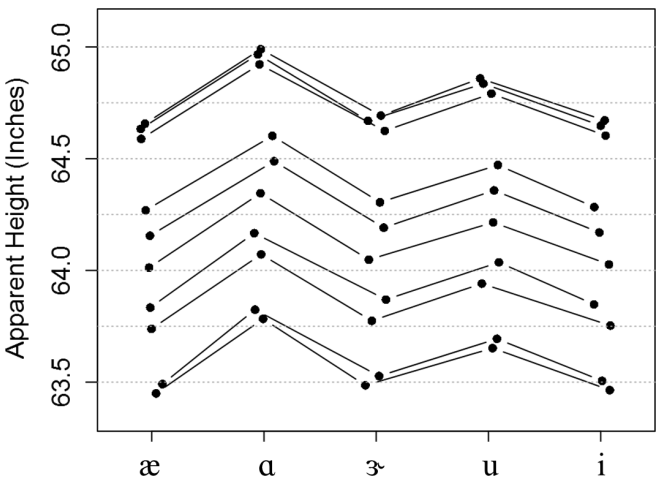

FIG. 7. Expected apparent-height differences for talkers in this experiment solely based on spectral information. Points indicate posterior means of average Vowel effects, added to the product of each talker's GMFF and the posterior mean GMFF slope effect.

this class of theories would only explain how listeners might recover something like a VTL estimate from individual tokens, and would need to be incorporated into a larger model of height perception that included roles for f0 and phoneme-specific spectral information.

The second class of theories that may explain VTL-based height perception are exemplar models (Goldinger, 1998; Johnson *et al.*, 1999), which suggest that phonemic representations are intrinsically associated with knowledge of the indexical characteristics of the talkers who produced them. As a result, all vowels produced by a given talker should be associated with roughly the same apparent-VTL and apparent-height. Although exemplar models can explain how listeners are able to use VTL information on a trial by trial basis, any approach to understanding perception of apparent-talker height that directly relies on experienced covariance patterns between talker height and speech acoustics is difficult to reconcile with some of the results presented here.

For example, it is not clear why some phonemes should consistently be associated with taller apparent talkers, since no such association exists in nature. Similarly, if apparent-height judgments strictly followed from experience, listeners should not rely so heavily on f0 for the estimation of adult female heights, given that the relationship in nature is weak and unreliable (this point is addressed further in Sec. IV C 2). As a result, the use of phoneme-specific spectral information and the reliance on f0 over VTL cues both suggest that listeners may exhibit a more general association between low frequencies and large apparent-talkers. This aspect of apparent-height perception suggests that height estimation involves at least some abstraction in the mapping of acoustics to height that does not directly follow from experience.

The final group of theories suggest that apparent-VTL estimates are available to listeners as a result of automatic processing carried out by the peripheral auditory system (Irino and Patterson, 2002; Ives *et al.*, 2005; Smith and Patterson, 2005; Turner *et al.*, 2006). It is argued that this processing provides phoneme-independent VTL estimates to listeners, and that it forms the basis of apparent-height perception in humans. Although these theories could explain how listeners estimate apparent-VTL, there are a few problems with any theory that posits easy access to phoneme-independent VTL estimates. First, listeners showed a great deal of variability in their use of VTL information, and some listeners barely respond to VTL differences at all in height estimation (Fig. 6). Second, height judgments themselves are not independent of phoneme-specific spectral information so that even if VTL and phonemic spectral information were segregated by the auditory system, they are re-conflated by listeners in estimating talker height. In light of these issues, estimation and reporting of phoneme-independent VTL information does not appear to be as easy, or consistent across listeners, as might be expected if VTL estimates were automatically available to listeners due to automatic auditory processing.

B. Phoneme biases in height judgments

Barreda (2016) reported that some vowel phonemes are consistently associated with taller or shorter apparent-talkers, independently of apparent VTL (i.e., phoneme-biases in height judgments). The results presented here confirm the presence of phoneme biases in height perception for absolute judgments collected for bVd words. Phoneme-biases appear to be largely predictable based on the spectral content of vowel sounds: The vowels associated with the largest apparent heights, /ɑ/ and /u/, are those with the lowest F1 and F2 values. The lack of an association between taller talkers and the rhoticized vowel /ɝ/, despite its very low F3, suggests that F1 and F2 differences more strongly determine phoneme biases than F3. For example, /u/ and /ɝ/ primarily differ in that /u/ has the lower F1 while /ɝ/ has the lower F3, but /u/ is consistently associated with taller talkers.

1. Magnitude of phoneme biases based on spectral information

The effects of phoneme biases reported here are smaller than those reported in Barreda (2016), which found that voices producing /ʊ/ tended to be identified as taller than those producing /æ/ even when /ʊ/ was presented with an 8% higher GMFF. In this experiment, the largest difference in apparent height between phonemes based solely on spectral information (0.33 in.) was equivalent to a difference of 3.2% in GMFF, and a 2.3% scaling difference in simulated VTL. The magnitudes of spectral phoneme-biases are compared to GMFF effects for the talkers in this experiment in Fig. 7.

Although the phoneme-bias effects are smaller than those reported in Barreda (2016), they are not small when compared to one standard deviation of GMFF in adult females, 0.04 log-Hz, representing variability of 4% in GMFF values (Hillenbrand *et al.*, 1995; Peterson and Barney, 1952). Hence, the largest spectral phoneme effect for the set of vowels used in this experiment is equivalent to 0.8 standard deviations in GMFF for adult females. Assuming adult female GMFF is approximately normally-distributed within-dialect, this means that if two adult females are selected at random, there is a 43% chance that the difference in their GMFF will be less than the largest spectral phoneme effect reported here. In other words, there is a 43% chance that a phoneme effect can re-order the relative apparent heights of any two adult female talkers with the same average f0. In light of this, the practical effects of spectrally-motivated phoneme biases on height perception appear to be non-trivial. As will be discussed in Sec. IV B 2, the effects of phoneme-biases are even larger if one considers predictable f0 variation between vowels.

The reduced magnitude of the phoneme-biases relative to those in Barreda (2016) may be explainable on the basis of some of the methodological differences between the two experiments. First, the stimuli in this experiment were bVd words in which the first and last consonant were stable for all talkers, while Barreda (2016) used isolated vowels. To the extent that voiced stop-consonants provide any information about talker height, they could work to stabilize estimates and diminish the phoneme-biases relative to isolated vowels. Second, Barreda (2016) used synthetic stimuli in which all source characteristics where matched across all stimuli. For the real voices used in the current experiment, f0

varied substantially between and within-talkers, in addition to many other idiosyncratic differences between the voices. The additional sources of variance may serve to diminish phoneme-biases either by overwhelming the relatively more-subtle spectral cues, or by allowing listeners to establish that some utterances were produced by the same apparent-talker, thereby suggesting that height should not vary across stimuli for that talker. Finally, the change from relative to absolute height judgments may fundamentally alter the process of determining apparent-talker height. For example, absolute estimation requires that listeners have some mapping from acoustics to a reportable absolute-value. In contrast, relative height judgments do not necessarily involve such a step since the sounds can be directly compared. Furthermore, the consideration of pairs of sounds presented in temporal proximity might involve local spectral-contrast effects in a way that would not arise when making absolute judgments (Assgari and Stilp, 2015; Stilp *et al.*, 2015).

2. Phoneme biases and predictable non-spectral variation

To this point phoneme-biases have been discussed as arising from predictable spectral variation in vowel phonemes. For example, /u/ will have a substantially lower F1 and F2 than /a/ in any given language, and so is likely to be associated with taller apparent-talkers, all other things being equal. However, it appears that phoneme biases in height perception may also arise as a result of the reliable covariation between f0 and different vowel phonemes.

The absolute f0 for a given vowel token will vary primarily due to inherent differences between talkers. However, f0 also varies systematically between vowel phonemes based on F1 (Whalen and Levitt, 1995). As a simple estimate of the regularity of within-subject log-f0 based on vowel category for the stimuli used in this experiment, a one-way ANOVA was run on centered log-f0 with vowel category as the only factor. This model indicates that 48% of within-talker variability in log-f0 is explainable on the basis of vowel category. A similar analysis carried out for the 48 adult female talkers in the Hillenbrand *et al.* (1995) data indicates that vowel category explains 33.6% of the variance in within-talker log-f0 for that set of talkers.

In combination with the predictable effect for within-talker variation in f0 outlined in Sec. III B 1, intrinsic f0 differences between phonemes have the potential to affect apparent-height judgments in a predictable manner. For example, based on the mean log-f0 for each phoneme across all talkers in the dataset (in Hz: æ = 193.5, i = 209.0, ɑ = 194.3, ɝ = 201.2, u = 204.5), and the posterior mean of the effect for centered log-f0, the following height differences are expected for each vowel phoneme, expressed in inches: æ = 0.134, i = −0.167, ɑ = 0.124, ɝ = −0.016, u = −0.080. These differences represent expected variability in the apparent height of a given talker attributable solely to expected variation in the intrinsic f0 of different vowel phonemes produced by that talker.

Effects for intrinsic-f0 will combine with intrinsic spectral characteristics to result in a net phoneme-bias which combines the two sources of information. As seen in Fig. 5, a low intrinsic pitch and low formant-frequencies combine in some vowels to give the impression of a taller speaker (/ɑ/), while for others low intrinsic formant-frequencies are counteracted somewhat by a high intrinsic pitch (/u/). The combined effects of intrinsic spectral and f0 information can also result in differing, and larger, patterns of phoneme biases. For example, the largest difference between phonemes based solely on spectral content was 0.33 in. between /æ/ and /ɑ/. However, the largest height difference between phonemes when f0 is also considered is between /i/ and /ɑ/ at 0.64 in. (seen in Fig. 5), which is equivalent to the effect of a 6% difference in GMFF between voices. A 6% difference in GMFF is 1.5 standard deviations of the variation in this cue for adult female talkers, meaning that when intrinsic-f0 is included in the consideration of phoneme biases, these biases are expected to overwhelm a large proportion of the variation in GMFF between adult female talkers.

3. Phoneme biases and size sound-symbolism

Size sound-symbolism is the association between size information and specific speech sounds (Hinton *et al.*, 2006; Ohala *et al.*, 1997). Several cross-linguistic studies have noted that high-front vowels are associated with morphemes denoting small sizes and low-back vowels are associated with large sizes more often than would be expected by chance alone (Haynie *et al.*, 2014; Ultan, 1978). In addition, listeners have intuitions about sound symbolic information in speech sounds. Shinohara and Kawahara (2010) presented L1 talkers of Chinese, English, Japanese, and Korean with disyllabic nonce words containing different vowels from an unknown "exotic" language with a rich lexical inventory of size adjectives. Listeners from all languages tended to associate high-front vowels with words denoting small sizes and low-back vowels with words denoting large sizes. This pattern of sound-size associations coincides with the findings reported here, where the largest overall difference in apparent-height is between a low-back vowel (/ɑ/) and a front-high vowel (/i/).

The cross-linguistic prevalence of associations between certain phonemes and specific semantic content, and the similar intuitions of talkers of diverse languages may be manifestations of phoneme-biases in apparent-talker height perception. When listeners are presented with unknown words and are asked to guess their size association, they may simply rely on apparent-talker height, which will be influenced by the specific acoustics of the speech sound. In the long-run, such tendencies may exert an influence on the phonological systems of different languages, resulting in convergent patterns of size sound-symbolism in different languages. If this is true, it would suggest that phoneme biases in height perception may have been hiding in plain sight in the form of common patterns of sound symbolism, and that these biases have enough perceptual salience to have practical effects on the phonological systems of different languages.

C. An updated approach to modeling apparent-talker height

Apparent height is typically modeled as a linear function of f0 and VTL information (e.g., GMFF), with a single

function used to estimate the height of all talkers by all listeners (Fitch, 1994; Smith and Patterson, 2005). Although such an approach works reasonably well as a first approximation, the results suggest that such models may not accurately reflect how listeners determine the apparent-height of talkers. The remainder of this section will outline suggested updates to the traditional approach to modeling apparent-height.

1. Class-conditional use of acoustic information

Perhaps the most serious problem with the idea that height perception is driven by a single estimation function for all talkers is that the use of acoustic information outlined in Sec. III will not generalize to female talkers of all ages.

Figure 8 presents the bivariate relationship between unstandardized height responses (in inches) and GMFF[5] in this experiment, and the average GMFFs and heights of females of different ages. As seen in Fig. 8, because of the high intercept and small changes in apparent height due to GMFF, listeners are expected to provide apparent-height estimates that are generally appropriate for adult females. In fact, the mean and standard deviation of unstandardized height responses across all listeners (64.2 and 2.79 in., respectively) mirror[6] the mean (64.2 in.) and standard deviation (2.81) of heights of females between 20 and 29 (Fryar *et al.*, 2012), and the distribution of heights of the talkers in this experiment (mean = 65.8, sd = 2.64). Although these characteristics result in good height estimates for adult female talkers, the same characteristics will hinder accurate height-estimation for younger females, which requires a substantially different relationship between GMFF and apparent height (dotted line, Fig. 8).

The non-generalizability of the adult-female model to the heights of younger females could be resolved if apparent-talker height were estimated on the basis of speech acoustics conditional on apparent-talker class. Such a process could be represented within a linear-model framework with a term for talker class (e.g., girl, boy, woman, man), as well as an interaction between this term and any acoustic cues whose interpretations vary based on class. Including class information in height-estimation models would allow listeners to (a) center their estimates around a reasonable average height for a class of talkers, and (b) use acoustic information in an appropriate manner given the variation between actual height and acoustics for talkers of the class. For example, listeners could switch from each of the lines presented in Fig. 8 based on whether they thought the talker was an adult or a child, and provide more reasonable height estimates in all cases.

Conditional height estimation would mean that in many cases talker class would be estimated from the same acoustic information used to determine talker height. Furthermore, class-conditional estimation of apparent-talker height would also allow for non-linguistic information such as instructions, or visual information, to affect the perception of apparent height. As a result, conditional height perception would be considerably more complicated than a mapping of acoustics to height without consideration of talker class. In any case, further research needs to be carried out regarding the potential involvement of talker-class information in the estimation of apparent-height.

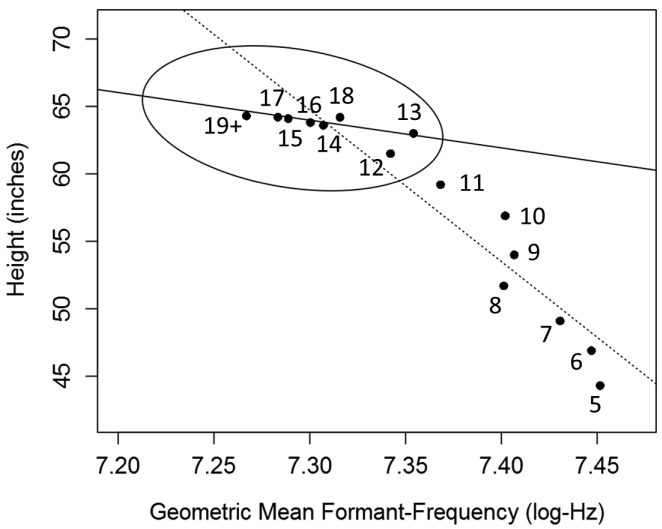

FIG. 8. Points indicate geometric mean formant-frequency (GMFF) for female age groups (Lee *et al.*, 1999), plotted against the average heights of females of the same ages (Fryar *et al.*, 2012). The ellipse encloses two standard deviations of the distribution of adult female heights and GMFF (Pisanski *et al.*, 2016), centered at mean GMFF and height values for women 15 and over. The solid line indicates the line of best fit for apparent (unstandardized) height and GMFFs in this experiment, while the dotted line shows the line of best fit for women from 5 to 17.

2. Non-empirical determination of the use of acoustic information

As noted in Sec. I, much of the research into apparent-height perception involves investigating the availability of height information in speech acoustics, and establishing the empirical relationships between characteristics such as GMFF and mean log-f0 with actual talker height in adults. However, several results in this experiment suggest that listeners may not necessarily base their estimates of talker height directly on the empirical relationships between acoustic variables and talker height in all cases.

For example, in isolation the presence of phoneme-biases in height perception may seem puzzling. For a listener to be able to distinguish /ɑ/ from /æ/, they must know that /ɑ/ will tend to have a lower F1 and F2 than /æ/, within-talker. As a result, we might imagine that listeners should correct for this sort of variation when estimating talker height, and not incorrectly associate phoneme-dependent formant differences with apparent-talkers of varying heights. Instead, listeners appear to respond to within-listener acoustic variability in a way that is not supported by the empirical relationship between acoustics and apparent height.

Similarly, in this experiment the magnitudes of GMFF and mean log-f0 effects were about equal. However, since variability in mean log-f0 is generally much larger than variability in GMFF between (and within) talkers in any given sample, the practical effect for changes in log-f0 will tend to be larger. However, VTL has been found to explain more of the variation in women's heights than f0, at 6.7% and 1.9%, respectively (Pisanski *et al.*, 2016). In addition, f0 is consciously manipulated by talkers quite easily, while VTL cues are substantially more difficult to change. A lifetime of

experience should have allowed listeners to establish that f0 provides substantially less reliable information about adult talker height than VTL, and that it can vary too much within-talker to be a reliable cue to talker height.

The inappropriate reliance on f0 differences when estimating apparent height, as opposed to using the more reliable VTL differences, is a common finding. For example, Rendall *et al.* (2007) report that f0 differences as small as 20 Hz can overwhelm contradictory VTL information in relative-height judgments. Similarly, Fitch (1994) found that a difference of 50 Hz (100 vs 150 Hz, a difference of 0.4 log-hz) has an effect on absolute height judgments that is five times as strong as a difference of 0.175 log-Hz in GMFF. For the sake of comparison, a difference of 50 Hz is small enough to be well within the normal variation in f0 for any given talker, while a difference of 0.175 log-Hz in GMFF is close to the difference in mean GMFF between adult males and children of either sex.

It is not being suggested that there is never an empirical foundation in the use of acoustic cues in the determination of apparent height. In fact, the use of GMFF to estimate the heights of adult females in this experiment is closely aligned with the empirical distribution of adult female heights and GMFFs, as seen in Fig. 8. Further, it is essential to determine the empirical relationships between speech acoustics and actual talker characteristics (as in Pisanski *et al.*, 2014; Pisanski *et al.*, 2016), as such research underlies the ability to make many of the claims contained in this work. However, it does not seem that the perception of apparent height will necessarily directly follow from the empirical relationships between speech acoustics and actual height. Instead, the use of acoustic cues in the estimation of talker height seems to vary in ways that may seem surprising *a priori*, and which may not be predictable on the basis of theoretical approaches to understanding the estimation of talker height. As a result, it seems that the effects of acoustic information on apparent-talker height should be empirically determined based on height judgments collected from listeners, rather than relying on the empirical relations between speech acoustics and actual talker-height.

3. Between listener variability

If apparent-height estimation were carried out using a single model common to all listeners, variability in the use of acoustic cues between listeners would only arise as a result of the underlying error that causes variation in apparent-height within-listener. However, the results presented in Sec. III C suggest that between-listener variation in the use of f0 and spectral information can be substantial, and may be an inherent characteristic of the way that different listeners estimate apparent-talker height. Furthermore, the variation in the responsiveness to VTL cues between-listeners is in line with previous reports that there are substantial individual differences in the ability to respond to apparent-VTL, that listeners can improve in reporting VTL after even brief periods of training, and that listeners with musical training exhibit an increased ability to report apparent-VTL (Barreda, 2016; Barreda and Nearey, 2013).

It is well known that the apparent-heights of different talkers will vary because of between-talker variation in VTL and f0. However, it appears as though there may also be substantial variation in the apparent heights of single talkers across multiple listeners as a result of the different relative weights that listeners give to VTL and f0 cues in height estimation. Consequently, between-listener variability in the mapping of acoustic information to apparent-talker height may add another layer of variation in the estimation of apparent-height from speech acoustics.

Between-talker variability in the use of VTL information in height estimation may reflect differing abilities to extract VTL information in a phoneme-independent manner across different listeners. It may also indicate that listeners differ in their tendency to rely on VTL when making talker-height judgements, irrespective of their ability to extract these estimates in a phoneme-independent manner. Although the root cause of this variation is not exactly clear, variability in the use of acoustic information in height estimation warrants further investigation. For example, it is not clear if these preferences are stable within-listener over time, or if the relative reliance on VTL information is associated with important physical, psychological, or social characteristics of different listeners. In any case, an understanding of the perception of apparent-talker height will likely include room for between-listener variation in the use of acoustic information that is distinct from error.

V. CONCLUSION

Listeners were asked to estimate the heights of ten adult-female talkers based on manipulated natural productions of five bVd words. Results show that some talkers were consistently identified as taller than others, indicating that speech contains information that listeners can use to estimate apparent-talker height with reasonable consistency from utterance to utterance. However, listeners exhibited phoneme biases in height perception due to inherent variation in spectral and f0 information across vowel phonemes. There was also substantial between-listener variation in the use of acoustic cues, resulting in predictable variation in the apparent-height of talkers, according to different listeners. Furthermore, although the determination of apparent-talker height is highly systematic, the use of acoustic cues may not necessarily align with the empirical relationship between speech acoustics and the actual heights of talkers. Finally, the determination of apparent-height may involve the use of information about apparent-talker class in a way that will tend to make height estimates more accurate, while also making the process of apparent-height estimation more complicated.

Overall, results suggest that apparent height estimation may involve a more nuanced use of acoustic information than is usually considered, with roles for both gross acoustic characteristics (e.g., GMFF) and token-specific information (phoneme-specific formant and f0). Furthermore, although results are consistent with VTL-estimation in height perception, the mechanism underlying this process is not exactly clear, especially given the lack of fit between the observed results, and predictions made by several prominent models of height (and speech) perception. Finally, further research is required on the

possible role for apparent talker class in the estimation of talker height, which would require a substantially more complicated estimation process, albeit one which would potentially result in interesting intersections between the determination of apparent height and many other salient indexical characteristics such as talker sex, age, and gender.

[1]As noted in Barreda (2016), inaccuracy in height judgments for adults likely stems from restricting height and speech acoustics to adult ranges. Such restrictions on ranges will weaken the correlation between any two variables, all other things being equal. Pisanski *et al.* (2016) report that vocal-tract length estimates explain 6.7% of the variation in female heights, meaning that the residual error in adult female heights after controlling for vocal-tract length will be 96.6% $[(1–0.067)^{1/2}]$ of the original error in heights, a modest reduction in error by any standard.

[2]Use of the term "vocal-tract length" in perception is meant to denote the average spectral scaling associated with a talker, which will be primarily determined by the vocal-tract length of that talker. Use of the term is not meant to suggest that listeners necessarily estimate the talker's vocal-tract length in units of length (e.g., inches).

[3]There was a weak correlation between actual talker height and average f0 ($r = -0.23$, 5% shared variance) and a moderate correlation between talker height and GMFF ($r = -0.51$, 26% shared variance). A linear regression model including GMFF and average f0 explains 26% of the variance in actual talker-heights for the stimulus voices.

[4]R^2 was calculated by finding the square of the correlation between predicted and estimated talker effects and those predicted using Eq. (5), for each of the posterior samples. This process results in a distribution than can be used to establish average and credible values for R^2.

[5]This comparison will focus on GMFF (i.e., VTL) information in height estimation because f0 explains very little of the variation in adult female heights (1.9%, Pisanski *et al.*, 2016), and f0 and GMFF averages are almost perfectly correlated for female talkers between the ages of 5 and 18 ($r = 0.94$; Lee *et al.*, 1999). Furthermore, the distribution of mean-log f0 for adult females overlaps almost entirely with f0s typical for much younger female talkers. As a result, the addition of f0 information to this illustrative example would not resolve any of the problems being discussed.

[6]The clustering of height estimates to adult female ranges cannot be explained by the ranges of the response ruler. First, the average reported height was 64 in. and the ruler midpoint was 63 in. If listeners wanted to report an overall lower average height (so as to be appropriate for younger women), but felt constrained by the range, the mean response should be relatively low on the ruler, not above the midpoint. Second, the standard deviation of responses (2.79 in.) was small relative to the range (18 in.). For example, the standard deviation of a uniform distribution between 54 and 72 (the ruler endpoints) is 5.2 in.

Assgari, A. A., and Stilp, C. E. (**2015**). "Talker information influences spectral contrast effects in speech categorization," J. Acoust. Soc. Am. **138**, 3023–3032.

Barreda, S. (**2016**). "Investigating the use of formant frequencies in listener judgments of talker size," J. Phon. **55**, 1–18.

Barreda, S., and Nearey, T. M. (**2013**). "Training listeners to report the acoustic correlate of formant-frequency scaling using synthetic voices," J. Acoust. Soc. Am. **133**, 1065–1077.

Boersma, P., and Weenink, D. (**2001**). "Praat, a system for doing phonetics by computer [computer program]," http://www.praat.org (Last viewed September 10, 2016).

Charlton, B. D., Taylor, A. M., and Reby, D. (**2013**). "Are men better than women at acoustic size judgments?," Biol. Lett. **9**, 20130270.

Collins, S. A. (**2000**). "Men's voices and women's choices," Anim. Behav. **60**, 773–780.

Fitch, W. T., and Giedd, J. (**1999**). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," J. Acoust. Soc. Am. **106**, 1511–1522.

Fitch, W. T. S. (**1994**). "Vocal tract length perception and the evolution of language," Brown University, Providence, Rhode Island.

Fryar, C. D., Gu, Q., and Ogden, C. L. (**2012**). "Anthropometric reference data for children and adults: United States, 2007–2010," National Center for Health Statistics. Vital Health Stat 11(252).

Gelman, A. (**2005**). "Analysis of variance—why it is more important than ever," Ann. Stat. **33**, 1–53.

Goldinger, S. D. (**1998**). "Echoes of echoes? An episodic theory of lexical access," Psychol. Rev. **105**, 251–279.

Haynie, H., Bowern, C., and LaPalombara, H. (**2014**). "Sound symbolism in the languages of australia," PLoS One **9**, e92852.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.

Hinton, L., Nichols, J., and Ohala, J. J. (**2006**). *Sound Symbolism* (Cambridge University Press, Cambridge, UK).

Irino, T., and Patterson, R. D. (**2002**). "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform," Speech Commun. **36**, 181–203.

Ives, D. T., Smith, D. R. R., and Patterson, R. D. (**2005**). "Discrimination of talker size from syllable phrases," J. Acoust. Soc. Am. **118**, 3816–3822.

Johnson, K., Strand, E. A., and D'Imperio, M. (**1999**). "Auditory–visual integration of talker gender in vowel perception," J. Phon. **27**, 359–384.

Joos, M. (**1948**). "Acoustic phonetics," Language **24**, 5–136.

Kruschke, J. K. (**2010**). "What to believe: Bayesian methods for data analysis," Trends Cogn. Sci. **14**, 293–300.

Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 98–104.

Lee, S., Potamianos, A., and Narayanan, S. (**1999**). "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," J. Acoust. Soc. Am. **105**, 1455–1468.

Nearey, T. M. (**1978**). *Phonetic Feature Systems for Vowels* (Indiana University Linguistics Club, Bloomington, IN).

Ohala, J. J., Hinton, L., and Nichols, J. (**1997**). "Sound symbolism," in *Proceedings of the 4th Seoul International Conference on Linguistics (SICOL)*, pp. 98–103.

Peterson, G. E., and Barney (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J. M., Röder, S., Andrews, P. W., Fink, B., DeBruine, L. M., Jones, B. C., and Feinberg, D. R. (**2014**). "Vocal indicators of body size in men and women: A meta-analysis," Anim. Behav. **95**, 89–99.

Pisanski, K., Jones, B. C., Fink, B., O'Connor, J. J., DeBruine, L. M., Röder, S., and Feinberg, D. R. (**2016**). "Voice parameters predict sex-specific body morphology in men and women," Anim. Behav. **112**, 13–22.

Pisanski, K., and Rendall, D. (**2011**). "The prioritization of voice fundamental frequency or formants in listeners' assessments of talker size, masculinity, and attractiveness," J. Acoust. Soc. Am. **129**, 2201–2212.

Plummer, M. (**2003**). "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling," in *Proceedings of the 3rd International Workshop Distributed Statistical Computing*, Technische Universitat Wien, Vienna, Austria, p. 125.

R Core Team (**2015**). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org (Last viewed June 30, 2016).

Rendall, D., Vokey, J. R., and Nemeth, C. (**2007**). "Lifting the curtain on the Wizard of Oz: Biased voice-based impressions of talker size," J. Exp. Psychol. Hum. Percept. Perform. **33**, 1208–1219.

Shinohara, K., and Kawahara, S. (**2010**). "A cross-linguistic study of sound symbolism: The images of size," in *Proceedings of the 36th Annual Meeting of the Berkeley Linguistics Society*, UC Berkeley, Berkeley, CA (February 6–7, 2010).

Smith, D. R. R., and Patterson, R. D. (**2005**). "The interaction of glottal-pulse rate and vocal-tract length in judgements of talker size, sex, and age," J. Acoust. Soc. Am. **118**, 3177–3186.

Stilp, C. E., Anderson, P. W., and Winn, M. B. (**2015**). "Predicting contrast effects following reliable spectral properties in speech perception," J. Acoust. Soc. Am. **137**, 3466–3476.

Turner, R. E., Al-Hames, M. A., Smith, D. R., Kawahara, H., Irino, T., and Patterson, R. D. (**2006**). "Vowel normalisation: Time-domain processing of the internal dynamics of speech," in *Dynamics of Speech Production and Perception*, edited by P. Divenyi (IOS Press, Amsterdam, the Netherlands).

Turner, R. E., Walters, T. C., Monaghan, J. J., and Patterson, R. D. (**2009**). "A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data," J. Acoust. Soc. Am. **125**(4), 2374–2386.

Ultan, R. (**1978**). "Size-sound symbolism," Univers. Hum. Lang. **2**, 525–568.

van Dommelen, W. A., and Moxness, B. H. (**1995**). "Acoustic parameters in speaker height and weight identification: Sex-specific behaviour," Lang. Speech **38**(3), 267–287.

Whalen, D. H., and Levitt, A. G. (**1995**). "The universality of intrinsic F0 of vowels," J. Phon. **23**, 349–366.