VOWEL NORMALIZATION AS PERCEPTUAL CONSTANCY

Santiago Barreda

University of California, Davis

This study investigates how listeners associate acoustically different vowels with a single linguistic vowel quality. Listeners were asked to identify vowel sounds as /æ/ or /ʌ/ and to indicate the size of the speaker that produced them. Results indicate that perceived vowel quality trades off with the perception of speaker size: different vowels can sound the same, and the same vowel can sound different when a different speaker is perceived. These findings suggest that vowel normalization is broadly similar to perceptual constancy in other domains, and that social, indexical, and linguistic information play an important role in determining even the most fundamental units of linguistic representation.*

Keywords: vowel normalization, speech perception, variation, vowel quality, sociolinguistics

1. Introduction. The phonemes that make up vowel systems are usually considered in terms of their perceptual qualities. For example, linguists may present the vowels of California English as in Figure 1a, where the axes represent the perceptual dimensions of vowel height (or openness) and vowel frontness (or backness). So, we may say that /u/ is 'fronted' in this dialect based on its position along the frontness axis, relative to its expected location. However, these representations do not provide any information about the acoustic realization of these vowels when produced by speakers of the dialect.

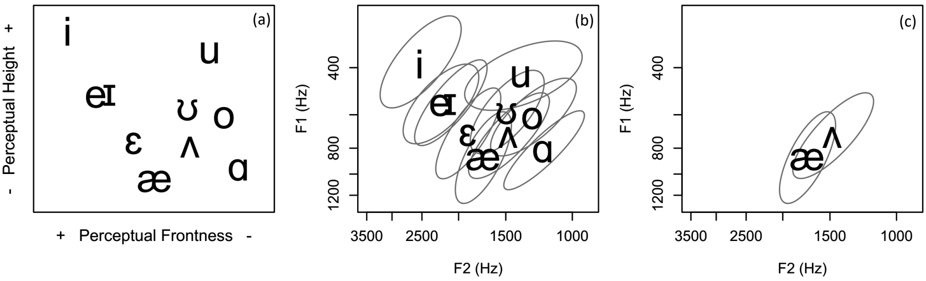

Figure 1. (a) The vowels of California English arranged on a space based on perceptual height and frontness. (b) Productions of thirty adult speakers of California English (fifteen women and fifteen men) in a database collected by the author. Ellipses enclose two standard deviations. (c) Productions of two ambiguous vowels are highlighted.

Vowel quality is the auditory sensation associated with vowel sounds, broadly similar to timbre. Perceived vowel quality is most strongly determined by the frequencies of the first two formants (Peterson 1961, Rakerd & Verbrugge 1985), with a weaker role for the third formant for some contrasts (Nearey 1989). However, between-speaker variation can result in ambiguity in the relationship between acoustics and vowel quality. For example, there are several locations in Figure 1b that could be associated with different vowel categories, as well as a range of locations that can be associated with each individual category. Despite this potential ambiguity, listeners can correctly identify vowel sounds at very high rates even when the sounds are presented in isolated /hVd/ syllables produced by unknown, randomly varying speakers (95% accuracy in

* I would like to thank Noah Silbert and Kristin Predeck for helpful comments on an earlier version of this manuscript. I would also like to thank Terry Nearey for his comments, and for the countless conversations that helped form and develop the ideas presented here.

Hillenbrand et al. 1995). This behavior has led to the suggestion that listeners carry out a process of perceptual 'normalization' that allows acoustically different sounds to be associated with 'the same' vowel quality. In fact, this perceptual ability is what allows researchers to create vowel-system organizations as in Fig. 1a, even though the sounds they use to arrive at such representations exhibit the variability seen in Fig.1b.

The term 'normalization' can connote different processes and outcomes based on the theoretical approach adopted (for a review, see Johnson 2005). Here, the term 'normalization' is used to refer to the perceptual process that determines vowel quality from speech acoustics, including the ability to associate acoustically dissimilar vowels with a similar perceived vowel quality. Developing an understanding of perceptual normalization is important to linguistics for theoretical and practical reasons. The operations involved in normalization will determine which between-speaker differences are 'normalizable' in perception. Variation in vowel sounds that is normalizable may be 'phone-preserving' to listeners (Nearey 1983), meaning that such differences will not affect perceived vowel quality and thus cannot be used to signal linguistic contrasts. By contrast, between-speaker variation that is not normalizable may be associated with differences in perceived vowel quality and can be used to communicate linguistic and social information between speakers. Thus, vowel normalization in perception influences and constrains both changes in vowel systems across time and phonologically and socially conditioned variation in vowel quality. As a result, a better understanding of perceptual normalization will benefit areas of linguistics 'downstream' from phonetics that relate to linguistic units specified in terms of vowel quality (e.g. phonemes and allophones).

In the pages that follow, I argue that vowel normalization should be considered an instance of perceptual constancy: the ability to perceive objects as having constant perceptual qualities despite differences in their associated sensory inputs. Although this comparison appears sporadically in the literature (e.g. Broadbent, Ladefoged, & Lawrence 1956, Kuhl 1979, Zhang, Peng, & Wang 2012), the full implications of this similarity have not previously been considered. Specifically, vowel normalization is similar to size constancy in the visual domain. I suggest that, in both cases, constancy is achieved by perceptual transformations that rely on the integration of multimodal sensory information, as well as social and cultural knowledge. Further, I argue that vowel normalization is guided by active cognitive control on the part of the listener, rather than representing the passive reception of speech sounds. Considering vowel normalization in this way suggests an ecological basis for the behavior and potentially imposes useful limits on the sort of variation that can be phone-preserving, which may aid the acquisition and communication of linguistic contrasts. Finally, the consideration of vowel normalization as perceptual constancy represents the integration of established findings from several different research domains and can naturally explain a seemingly disparate set of observed phenomena in the literature on speech perception.

1.1. Phone-preserving variation in formant patterns. The uniform scaling hypothesis (also known as the constant ratio hypothesis) is one of the earliest attempts to explain phone-preserving variation in formant patterns, its general outlines being suggested at least as far back as Lloyd 1890. The uniform scaling hypothesis states that the perceived quality of vowel sounds is more likely to be maintained when variation in formant patterns is according to a single scaling parameter, relative to uncoordinated changes to individual formant frequencies. This hypothesis has substantial empirical and theoretical support in the perception of vowel sounds. As noted by Miller (1989) and Johnson (2005), many seemingly different models of vowel perception are variations of the general insight that formant patterns with the same vowel quality tend to be relatable by a single scale factor (e.g. Peterson 1961, Sussman 1986, Patterson & Irino 2014). Perceptual experiments indicate that uniform scaling of formant patterns preserves perceived vowel quality in a way that independent scaling of the formants does not (Nearey 1983). Furthermore, vowel quality can be maintained under uniform scaling for up to 200% differences in formant frequencies (Smith et al. 2005, Assmann & Nearey 2008). Because of this, uniform scaling of formant patterns is frequently used to simulate speakers of different sizes in experimental work on size perception (e.g. Smith & Patterson 2005) and in the entertainment industry (e.g. Alvin and the Chipmunks; Winer 2012), an approach that would not be possible if such shifts resulted in changes in the phonemic identity of speech sounds. In contrast, relatively small independent changes to individual formant frequencies can be used to signal differences in vowel quality.[1]

Variation of formant patterns according to a single spectral-scaling parameter (i.e. uniform scaling) results in spectra that differ in the expansion or contraction of a fixed pattern along the frequency axis, as in Figure 2a. Another way to look at this is that a single vowel quality can be associated with spectra differing in their ‘size’ expansion along the frequency axis, so long as their general, underlying spectral ‘shape’ is preserved (Irino & Patterson 2002). Changing the length of a vocal tract (or any other tube resonator) by applying equal proportional changes to all dimensions (Figure 2b) will result in output spectra that differ according to a single spectral-scaling parameter related to the proportional difference in length between the vocal tracts (Wakita 1977). For example, increasing the length of a vocal tract by a factor of 1.1/1 (110% of the original length) will result in a decrease in output formant frequencies by a factor of 1/1.1 (90.1% of original values). Thus, the spectral variation seen in Fig. 2a is of the sort expected when speakers who differ strictly in vocal-tract length adopt similar gestures (as in Fig. 2b). Variation according to uniform scaling will result in phonemes that vary along lines parallel to log F1 = log F2, and in vowel spaces that differ between-speaker as translations along these lines, as in Figure 2c (Nearey 1978). The spectral-scaling parameter associated with differences in vocal-tract length (or resonator size more generally) is referred to as ψ, where speakers with longer vocal tracts have lower values of ψ and produce lower formant frequencies overall.

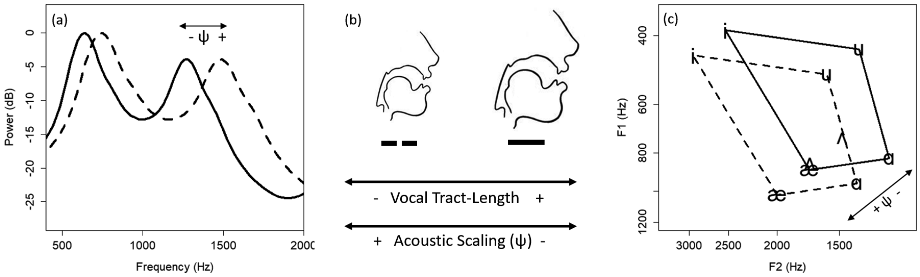

Figure 2. (a) Vowel spectra differing in their spectral scaling (ψ), and formant frequencies, by 16%. Line types indicate corresponding vocal tracts and vowel systems in b–c. (b) Speakers differing in vocal-tract length and acoustic scaling ψ. (c) Vowel spaces corresponding to the speakers presented in (a). Variation strictly according to ψ constrains vowel spaces to differ along lines parallel to log F1 = log F2, indicated by the direction of the arrows.

[1] A clear example of this comes from experiments where listeners classify vowels varying along formant continua (e.g. Johnson 1990, Glidden & Assmann 2004). There can often be substantial differences in perceived vowel quality between adjacent steps in the continua, despite small differences in formant frequencies between steps (e.g. 3% for F1, 1% for F2 in Johnson 1990).

We may formalize the relationship between the underlying spectral shape, the spectral-scaling parameter, and the acoustic formant pattern as in equation 1. Consider an underlying spectral shape, V_v, associated with some vowel quality v. The formant pattern produced by speaker s for this vowel, F_{vs}, can be thought of as the product of the underlying shape (V_v) and the speaker-specific spectral-scaling parameter ψ_s. For example, increasing the value of ψ_s in equation 1 will result in an expansion of the underlying spectral pattern (as in Fig. 2a), resulting in higher formant frequencies based on the same underlying spectral shape. Several different parametrizations of ψ have been proposed through the years, including the geometric mean formant frequency (Nearey 1978), the arithmetic mean formant frequency (Pisanski & Rendall 2011), and the average spacing between formants (ΔF; Reby & McComb 2003a). Usually, the different parametrizations of ψ provide estimates that are strongly correlated, and in many cases in the absence of error they would be perfectly correlated.[2] Here, between-speaker variation in ψ is discussed directly in terms of proportional differences in expected formant frequencies, so if the ψ_s of one speaker is 10% higher than that of another, the first speaker is expected to produce formants that are 10% higher overall than those of the second speaker for any given vowel phoneme.

(1) $F_{vs} = V_v \times \psi_s$

The uniform scaling hypothesis does not imply that all between-speaker variation in formant patterns for a phoneme should be according to uniform scaling, but rather that these differences are most likely to preserve vowel quality. For example, speakers of California English differ in terms of their participation in different aspects of the California Vowel Shift. Podesva et al. (2015) suggest that these differences are used by speakers to transmit information about identity and ideology. Differences in the production of individual phonemes between speakers will result in within-phoneme variation that is not according to uniform scaling. For example, differences in /u/-fronting mean that two speakers who produce approximately the same F1 and F3 frequencies for /u/ may differ in their F2 in a somewhat unpredictable manner based on their specific amount of fronting. Effectively, this means that different speakers can have subtly different specifications of the underlying spectral pattern (V_v) for the ‘same’ phoneme. The uniform scaling hypothesis does not suggest that such variation should not exist, but rather that this sort of variation is likely to be ‘heard’ by listeners, thereby allowing the transmission of linguistic and social information between speakers.

1.2. Spectral scaling and apparent-speaker characteristics. Since vocal-tract length is strongly correlated with speaker height across the entire human population (Fitch & Giedd 1999), and spectral scaling (ψ) varies predictably with vocal-tract length, ψ provides useful information about speaker height and other correlated characteristics such as age and sex. Listeners are aware of this covariation and use information related to ψ to estimate the height (Smith & Patterson 2005), age (Barreda & Assmann 2018),

[2] Consider a speaker who produces formant frequencies that are exactly 10% higher than a second speaker, for every formant and vowel. The first speaker’s geometric mean formant frequency will be exactly 10% higher than that of the second. Further, their ΔF must also be exactly 10% higher since the distance between all formants has increased by 10%. As a result of this tendency, selection of a specific parametrization of ψ most suitable for an objective depends on the goals of the research and the statistical properties of the different estimators. It is also important to consider that estimates of ψ are contingent on the specific sample (and language) used to estimate them, complicating the estimation (and definition) of the ‘true’ ψ for a speaker (Barreda & Nearey 2018). For example, estimates of ψ for a single bilingual speaker would very likely differ across their two languages because of inventory differences between the languages, despite a single underlying anatomy.

and sex of speakers (Smith, Walters, & Patterson 2007, Hillenbrand & Clark 2009), consistently associating lower apparent values of ψ with taller apparent speakers. A diverse set of mammal species such as red deer (Reby et al. 2005), koalas (Charlton et al. 2012), and dogs (Taylor, Reby, & McComb 2010) also exhibit behavior consistent with the use of ψ to assess body size. This suggests that the association between ψ and perceived size may be a fundamental aspect of auditory perception in mammals, and that this association likely predates speech communication in humans by a wide margin (Irino & Patterson 2002).

It has been noted that the use of size information in communication involves processes broadly similar to normalization in vowel perception (Barreda 2017). In fact, when viewed from a certain perspective, other mammal species also exhibit behavior that resembles normalization. For example, red deer produce several different types of calls, each of which serves a different communicative function (Reby & McComb 2003b). As a part of their mating behavior, male red deer produce a 'common roar' (Reby & Charlton 2012), which 'is characterized by an initial drop in formant frequencies, which then plateau at minimum values before they rise again at the end of the roar' (Charlton, Reby, & McComb 2008:2937). In red deer, size is strongly correlated with neck length, so differences in deer size directly relate to differences in ψ in the calls of red deer (Reby & McComb 2003a). Female red deer prefer males whose roars suggest a lower ψ (Charlton, Reby, & McComb 2007), and male deer will increase their roaring rates and lower their ψ when exposed to roars from a male deer with a low apparent ψ (Reby et al. 2005). These behaviors suggest that the formant structure of roars provides 'individuals with the means to assess the [vocal-tract length] and probable body size' (Reby et al. 2005:946) of other deer. So, two roars whose formant patterns suggest differences in ψ are interpreted by deer as signaling differences in deer size, rather than reflecting a change in call type—they are interpreted as two instances of the 'same' call. This behavior is broadly analogous to a listener who can identify different realizations of a vowel phoneme as being 'the same' despite differences in their formant patterns related to ψ. To complete the analogy, in such a situation the human listener is likely to make the same association as the deer and use the differences in formant patterns to infer differences in speaker size.

Although not typically considered in this manner, the uniform scaling hypothesis effectively suggests that between-speaker variation is most likely to be phone-preserving in cases where it can be interpreted as signaling a change in speaker size or other indexical characteristics dependent on ψ. In fact, this may be precisely why uniform scaling is phone-preserving: because this sort of variability 'sounds' like differences in size to human listeners, it is interpretable as nonlinguistic between-speaker variation. Thus, when speakers produce formant patterns that differ primarily according to uniform scaling, this variation will 'feel' to listeners like a change in the size of the speaker, rather than a change in the linguistic content of the signal. As a result, between-speaker variation according to uniform scaling potentially allows vowel normalization to operate via mechanisms already in place to interpret size information from sounds (Patterson & Irino 2014).

1.3. Active participation in the maintenance of uniform scaling. If vocal tracts differed only in length and speakers adopted similar gestures in production, they would produce formant patterns that differed strictly according to a single spectral-scaling parameter (ψ). However, it has long been noted that speakers can vary in their vocal-tract geometry in significant ways above and beyond differences in length. It has been suggested that anatomical differences between the vocal tracts of adult males and females should result in nonuniform scaling of formant patterns for different speakers (Fant 1966, 1975). In fact, between-speaker differences in vocal-tract anatomy are pervasive, with substantial differences in the geometry of the vocal tract as a function of age, height, and sex (Story et al. 2018). This means that not only are there significant anatomical differences between adult males and females, but that such differences also exist between all speakers, and that all speakers will pass through a wide range of different vocal-tract geometries as they age into adulthood. Thus, if deviations from uniform scaling followed strictly from anatomical differences between speakers, we should expect to see large, consistent deviations from uniform scaling in the formant patterns produced by the speakers of a dialect for a given phoneme as a function of age, sex, and height. This does not appear to be the case, and speakers tend to produce formant patterns that largely conform to uniform scaling (Turner et al. 2009, Barreda & Nearey 2013).

It has been suggested that speakers actively adopt compensatory gestures in order to maintain variation in formant patterns according to uniform scaling, despite anatomical differences (Nordström & Lindblom 1975, Turner et al. 2009). In fact, since speakers necessarily pass through many vocal-tract geometries as they grow from children to adults, the only way to produce consistent formant patterns at different ages would be to update gestures to match changing anatomies. This would require speakers to continuously monitor their acoustic outputs and update their gestures as necessary as their vocal-tract structure changes through time. The monitoring of speech output to maintain an expected formant pattern is behavior that has been observed in several different experimental paradigms. For example, bite-block experiments show that when jaw movement is impeded, speakers can drastically alter their articulation in order to produce formant patterns similar to their unimpeded productions (Fowler & Turvey 1980). Similarly, in experiments where speakers receive altered auditory feedback regarding their own productions, speakers immediately adopt compensatory gestures to try to maintain expected formant patterns given the altered feedback (Purcell & Munhall 2006). Although these compensatory mechanisms have been demonstrated at single points in time, this monitoring could also be employed to maintain a constant vowel quality in the face of age-related changes in vocal-tract anatomy. In the absence of such behavior, speakers would rigidly recreate the gesture they learned as young children in their adult bodies, resulting in large deviations from the formant patterns they initially produced as children.

In this view, the maintenance of uniform scaling of formant patterns in a speech community requires effort on the part of speakers and should not be thought of as occurring by default. Although such behavior results in more work for speakers, the maintenance of uniform scaling may facilitate perception by reducing variation in formant patterns to the ‘simple’ case of variation strictly according to vocal-tract length. When this occurs, between-speaker variation will typically be of the kind that tends to be treated as linguistically irrelevant by listeners (i.e. ‘normalizable’ size variation) rather than reflecting unpredictable anatomical differences between speakers.

1.4. Ambiguity in spectral patterns. Variation between speakers can result in between-category overlap in the formant space, resulting in formant patterns that are ambiguous with respect to phonemic category. According to the uniform scaling hypothesis, vowel phonemes that differ along the dimension associated with differences in ψ (e.g. /æ/ and /ʌ/ in Fig. 2c) in a formant space may be most perceptually confusable to listeners. This is because the differences in the formant patterns between these phonemes also resemble differences between speakers according to spectral scaling, with one vowel having an inherently 'bigger' or more expanded formant pattern (/æ/ in this case). For example, the spectra presented in Fig. 2a could represent the productions of /ʌ/ by the two speakers in Fig. 2b, or they could represent the productions of /æ/ (dashed line) and /ʌ/ (solid line) by the smaller speaker. However, even in the absence of ambiguity about phonemic categorization, there may always be PHONETIC ambiguity for any given vowel token, even within a single category. This is because variation along the axis indicated in Figure 1c above can represent a difference in vowel quality (and the underlying spectral pattern V_v), a difference in ψ, or a combination of the two. As a result, two tokens of a single phoneme that vary along one of these lines either can indicate subphonemic differences between speakers with the same ψ, the same quality for speakers of different ψ, or some combination thereof.

A comparison can be made between the spectral ambiguity resulting from variation in ψ and the ambiguity that results from the distance-related scaling of objects in the visual domain. The perceived visual size of an object is related to the size of the retinal image in the eye of the observer. All other things being equal, an object with a larger retinal image than another will be perceived as larger. However, as distance from the observer decreases/increases, the retinal image is scaled proportionally and uniformly across every dimension (i.e. height-scale factor = width-scale factor). This means that, for example, as a square object moves further away from or closer to an observer it may appear to be a larger or smaller square, but not a rectangle. Consider two objects with roughly the same shape that differ primarily in their intrinsic scaling, such as a car (the inherently 'big' pattern, /æ/) and a one-eighteenth scale replica of the car (the inherently 'small' pattern, /ʌ/). When observing images of these objects in isolation, as in Figure 3a–b, it is difficult to know if the apparent sizes of the objects reflect differences in their inherent sizes, differences in the distance-related scaling of the objects, or some combination of the two. Just as in the case of /æ/ and /ʌ/, the shapes in Fig. 3a–b can be ambiguous to observers because the difference in their intrinsic shapes is of the same kind as the orderly scaling variation that exists in the domain. Note also that when the image is scaled nonuniformly (Figure 3d), the result is no longer the perception of changing size or distance, but rather the impression that the object has changed.

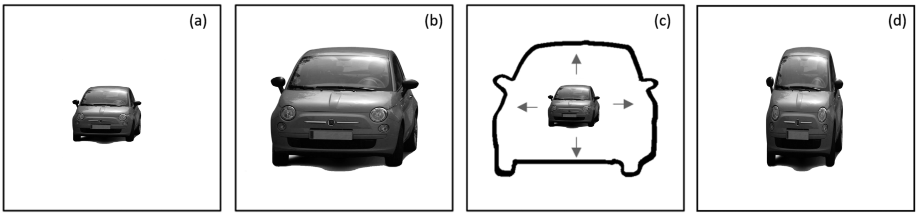

FIGURE 3. (a)–(b) Images of cars differing in size, distance, or some combination of the two. (c) An example of uniform scaling of a visual pattern, the kind of scaling expected according to the distance of visual objects. (d) An example of nonuniform scaling of the image, resulting in an apparent change in the characteristics of the object.

Representations such as those in Fig. 1b overstate the confusability of formant patterns by omitting information in the higher formants (the formants above F2). The higher formants may help disambiguate many cases of between-category overlap seen in the two-formant space in at least two ways. First, they may provide more reliable information about ψ (and vocal-tract length) for a speaker by being relatively stable across vowel categories within-speaker, relative to the lower formants (Wakita 1977, Lammert & Narayanan 2015). Second, the higher formants may reduce ambiguity in the lower formants by directly specifying the spectral pattern to a greater extent. For example, front vowels tend to have F3 and F4 values that are 200–400 Hz higher than those of back vowels (Hillenbrand et al. 1995:3103, table V), meaning that some ambiguity seen in Fig. 1b could potentially be resolved by considering the position of the higher formants. However, enough ambiguity remains that there is a many-to-many mapping between spectral shapes and perceived vowel quality: one spectral shape can be associated with different vowel qualities, and different spectral shapes can be associated with similar vowel qualities. As noted by Nusbaum and Magnuson (1997), many-to-many mapping problems are inherently nondeterministic and cannot typically be solved by cognitively passive (open-loop) perceptual processes. Passive processes are characterized by static relationships between inputs and outputs, resulting in predictable perceptions given fixed stimulus properties regardless of the listening situation (e.g. Syrdal & Gopal 1986, Miller 1989). However, many-to-many perceptual problems require variable mappings between inputs and outputs. As a result, Nusbaum and Magnuson suggest that many-to-many mapping problems, including vowel normalization, need to be solved with cognitively active (closed-loop) perceptual processes, which can relate sensory information to perceptual units in a context-dependent manner, often relying on feedback from other sensory or cognitive processes.

1.5. Disambiguation through the consideration of spectral scaling. It has long been noted that potentially ambiguous formant patterns relate to perceived vowel quality based on the relative location of the vowel within the speaker's vowel system (Joos 1948, Ladefoged & Broadbent 1957, Nearey 1989). This suggests that a vowel at the location indicated by the circle in Figure 4a would most likely be identified as an /ʌ/ for the speaker in Figure 4b, and as an /æ/ for the speaker in Figure 4c. So, although the location indicated in the figure is associated with a fixed formant pattern, it is potentially associated with two different perceived vowel qualities based on their position in the speaker-dependent vowel space (Figure 4d). This ambiguity can be expressed in terms of the relation presented in equation 1 above: in Fig. 4a, a single acoustic formant pattern can potentially reflect two underlying spectral shapes as a result of offsetting differences in spectral scaling.[3] This suggests that in some cases, ambiguity in formant patterns may be resolved by controlling for the spectral scaling associated with the speaker.

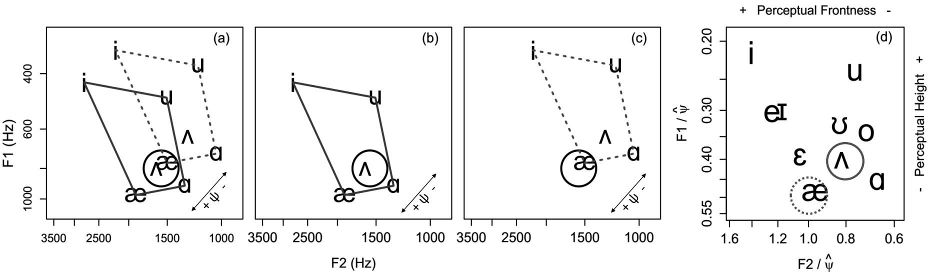

Figure 4. (a) A formant pattern falling in a potentially ambiguous location in the vowel space. (b)–(c) Two different interpretations of the vowel based on differing estimates of ψ. (d) Each ψ estimate is associated with a different relative (ψ-adjusted) location in the vowel space, resulting in different possible vowel qualities for the single formant pattern. Relative locations were found by dividing the ambiguous formant pattern in (a) by the geometric mean formant frequency of the vowels of the speakers in (b) and (c), as in equation 3. Line types indicate associations between vowel spaces and interpretations.

[3] I highlight situations where there is phonemic ambiguity because these cases are easiest to consider, and they are potentially the most linguistically interesting. However, as noted earlier, there is potentially continuous phonetic ambiguity for any formant pattern, meaning that there are many more than two potential interpretations of the location indicated in Fig. 4a.

The relationship presented in equation 1 can be rearranged to isolate the underlying spectral pattern (V_v) from variability related to spectral-scaling differences between speakers, as in equation 2. The inversion in equation 2 follows directly from the relationship between resonator size and formant patterns presented in equation 1: if formant patterns (F_{vs}) are the product of some underlying spectral shape (V_v) and ψ_s, then dividing the formant pattern by ψ_s should reveal the underlying spectral shape. In practice, listeners only have access to estimates of ψ_s, $\hat{\psi}_s$, resulting in estimates of the underlying pattern $\hat{V}_v$ as in equation 3. Dividing the observed formant pattern by an estimate of ψ_s for a speaker results in a ψ-adjusted representation that controls for differences in scaling between speakers (as in Fig. 4d). By controlling for differences in spectral scaling, these ψ-adjusted formant frequencies relate more directly to perceived vowel quality than the acoustic formant pattern does.[4] In the example provided in Fig. 4, the ambiguous location in the formant space could be classified unambiguously, given a known estimate of $\hat{\psi}_s$ for the speaker.

(2) $V_v = F_{vs} / \psi_s$

(3) $\hat{V}_v = F_{vs} / \hat{\psi}_s$

Although controlling for ψ in vowel perception may help resolve ambiguity in formant patterns, this also results in a circular 'chicken and egg' problem where the determination of vowel quality relies on knowing ψ, and knowing ψ relies on the determination of vowel quality. In Fig. 4, we see that in order to identify the ambiguous sound the listener must commit to an estimate of ψ for the speaker (Nearey 1978, Turner et al. 2009, Patterson & Irino 2014). However, estimating ψ accurately for a speaker requires that vowel quality be known: the point indicated in Fig. 4a is associated with different ψ estimates, and different apparent-speaker characteristics such as height, based on the vowel quality (and phonemic category) assumed for the sound.

Similar perceptual problems arise in other domains. As a result of the relationship between retinal image size, linear size, and distance-related scaling, an observer must estimate the distance of objects in order to estimate their linear size. This results in a trade-off between the apparent distance of an object and its apparent linear size (Holway & Boring 1941). For example, imagine a situation where the images of cars of different sizes were presented at random, without any contextual visual information, as in Fig. 3a–b. Suppose that observers were asked to determine: (i) whether the image represented a full-sized car or a one-eighteenth scale model of the car, and (ii) how far the car is from the photographer. We would expect listeners to identify the image as a full-sized car that is far from the observer, or a one-eighteenth scale replica that is close, but not vice versa. This presents a parallel situation to the ambiguous vowel in Fig. 4, which can only be either an /æ/ produced by a taller person, or an /ʌ/ produced by a shorter person. In both cases, decisions regarding the inherent shape (including the inherent 'size') of an object must be made together with decisions about the scaling that has been applied to the shape. Furthermore, just as in the identification of vowel sounds, ambiguity in visual objects is reduced by presenting the objects at a fixed distance, thereby equating the distance-related scaling applied to them. Presenting visual objects in such a 'distance-adjusted' condition would be analogous to the operation described in equation 3, and the representation shown in Fig. 4d.

[4] In fact, this approach is taken by several algorithmic normalization methods proposed for use in quantitative sociolinguistic investigations (e.g. Nordström & Lindblom 1975, Nearey 1978, Johnson 2018). Although these methods differ in their parametrizations of ψ, they all suggest that vowel quality is best reflected by considering ψ-adjusted, rather than absolute, formant patterns.

1.6. Vowel normalization as perceptual constancy. The problem of lack of invariance is not unique to speech but is instead ubiquitous in perception. For example, as a result of differences in distance, lighting, rotation, and so on, any known visual object will rarely look the same twice. Perceptual constancy is the tendency to perceive objects as having consistent perceptual qualities despite varying associated sensory inputs. For instance, the intensity of the light reflected from the white area of a sheet of paper under indoor lighting is roughly equal to the intensity reflecting from the black print on the same sheet when viewed under sunlight (Kaiser & Boynton 1996:199). Despite this, the paper appears white and the print appears black in both situations. Generally speaking, perceptual constancy tends to involve the reinterpretation of an apparent change in the characteristics of the object as information about the context in which the object is being observed. So, differences in the hue or brightness of an object might be interpreted as differences in lighting conditions (Hilbert 2005), and differences in the apparent shape of an object can be interpreted as differences in its rotation with respect to the observer (Slater & Morison 1985). As a result, maintaining perceptual constancy involves being able to solve the sorts of 'circular' problems posed by normalization, where determining the properties of an object and determining the properties of the environment depend on each other.

Vowel normalization is broadly similar to size constancy in the visual domain. Although the size of the retinal image changes as a function of distance, humans do not tend to perceive objects moving away from them as decreasing in size. Instead, the change in the retinal image size is interpreted by the observer as a change in object distance (Gogel 1969). Just as in vowel normalization (e.g. equation 3), subjective size constancy in the visual domain involves a 'cognitive scaling operation' (Sperandio & Chouinard 2015:253) whereby an apparent change in stimulus properties is reinterpreted as contextual information. Because of this process, visual objects maintain a constant apparent size and shape (and therefore identity), even as their representation in the eye of the observer changes.

In the case of vowel normalization, the 'objects' being held constant are phonemes of a dialect. By definition, each phoneme has a consistent (or nearly consistent) vowel quality across speakers of the dialect. In some cases, differences in the acoustic realization of vowel sounds are reinterpreted by listeners not as changes to the quality of the vowel (and therefore a change in the properties of the 'object'), but rather as indicating a change in the 'environment' in which the vowel was produced, in this case the speaker's vocal tract. This account of speech perception suggests that orderly variation according to speaker indexical characteristics is not 'noise' the perceptual system must filter away, but rather a rich source of information that can potentially facilitate speech perception. Here, it is suggested that between-speaker variation (within dialect) according to uniform scaling of spectral patterns is reinterpreted as variation in ψ, which also informs the perception of indexical and social information regarding the speaker. In the same way, the cars in Fig. 3a–b, and visual objects more generally, are simultaneously recognized as differing while also representing instances of a single object with fixed properties.

1.7. Apparent-speaker characteristics and the disambiguation of spectral patterns. Identifying vowel sounds requires committing to an estimate of vowel-space location (and ψ_s) for the speaker. As a result, in vowel perception listeners are effectively tasked with selecting the most likely ψ_s with which to interpret the sound (Nearey 1978). How listeners accomplish this for unknown speakers is a subject of debate. Although human listeners do use prior context to help identify speech sounds (Ladefoged & Broadbent 1957), perceptual accuracy is still very high even for isolated syllables presented from randomly varying speakers. This indicates that although prior context is useful, it is not strictly necessary for accurate perception. Considering the relationship between vowel normalization and size perception may help us understand how listeners resolve this problem. Since ψ is strongly related to highly salient indexical characteristics such as sex, height, and age, there may be a wealth of information (both auditory and nonauditory) available to help disambiguate such situations by suggesting constraints on plausible ranges of ψ for a speaker. Thus, information about the speaker may affect perceived vowel quality 'indirectly' (Johnson 1990) by informing the estimate of ψ used to interpret the vowel sound (Nearey & Assmann 2007). In this way, ψ_s can be estimated accurately from a single vowel token, potentially solving the long-standing problem of accurate perception in the absence of context (Nearey & Assmann 2007).

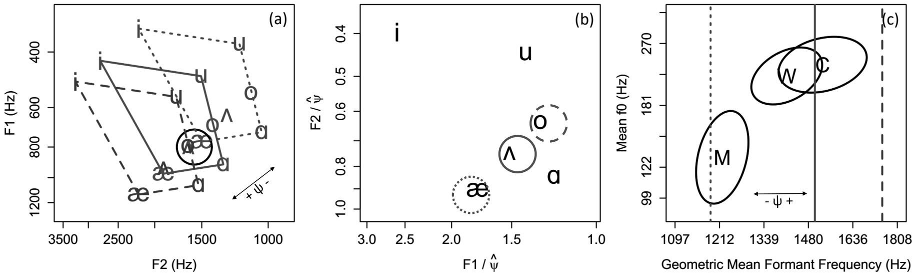

Figure 5. (a) An ambiguous location is compared to three different vowel spaces, each associated with a different estimate of ψ. (b) The ambiguous pattern in (a) is adjusted for the ψ implied by each vowel space in (a) using equation 3. This reveals three possible interpretations of the underlying formant pattern given different estimates of ψ. Line types indicate associations between vowel spaces and interpretations. (c) Distribution of average mean f0 and geometric mean formant frequency (used to index differences in ψ) for forty-five men (M), forty-eight women (W), and forty-six children between ten and twelve years of age (C) from the Hillenbrand et al. (1995) data set. Ellipses enclose two standard deviations of speakers of each category. Vertical lines indicate values of ψ implied by the different vowel spaces in (a) and interpretations in (b). Each vowel space, and associated phonetic judgment, suggests a different sort of speaker.

An illustrative example of how this might work is presented in Figure 5, which shows an ambiguous lower-formant pattern and three possible interpretations.[5] Each interpretation is associated with a different estimate of ψ and different expectations about speaker f0, sex, age, and so on. In order for the listener to identify the ambiguous vowel as /æ/, they must decide on a low ψ (and a long vocal tract). Evidence to suggest that the speaker is tall (e.g. an f0 of 110 Hz, a male face, low higher formants) may suggest a low ψ, making a listener more likely to identify the sound as an instance of /æ/. If the vowel were instead presented with an f0 of 230 Hz, a listener may expect a smaller speaker (and shorter vocal tract) and therefore a higher ψ. In this case, the listener may be more likely to identify the ambiguous pattern as an instance of /ʌ/. Finally, we see that the ψ implied by the /o/ interpretation is very high, suggesting that the speaker should be small even relative to ten-year-old children. Listeners may be skeptical of this interpretation and unlikely to identify the vowel as an /o/ in the absence of concurring information (e.g. very high higher formants, visual information, etc.). As a result, what

[5] The model being described here is the sliding template model of vowel perception, first presented in Nearey 1978 and elaborated on in Nearey & Assmann 2007.

seems to be a circular problem for the listener may be resolved by using all available information to arrive at a holistic impression of a speech event, including a coherent integration of the perceived characteristics of the speaker and the utterance.

There is evidence suggesting that nonspectral information can affect perceived vowel quality indirectly in the manner outlined above. Johnson (1990) presents a series of experiments showing that the effect for f0 on perceived vowel quality is strongest when associated with perceived changes in speaker size. Johnson concludes that the results 'can be explained if we assume that hearers construct a representation of the talker in the process of auditory word recognition' (1990:652). Perceived vowel quality can also be affected by nonauditory information such as pictures of speakers of different sexes (Glidden & Assmann 2004) or even instructions about the sex of the speaker (Johnson, Strand, & D'Imperio 1999). In all of the aforementioned experiments, indirect effects on perceived vowel quality occurred in the direction predicted by the relationship between changes in ψ and perceived vowel quality: cues that suggest a larger speaker (e.g. a low f0, an adult male face) result in increases in the perception of open/low vowel quality for any given vowel sound. This is because when a lower ψ is assumed, any given formant frequency appears relatively higher, resulting in perceptually lower and more open vowels.

Such 'indirect' information can also play an important role in assessing the identity of objects in the visual domain. Figure 6a presents images of cars that have been modified according to uniform scaling. In the absence of any contextual information, these cars appear to vary in size, but the image is ambiguous. However, converging lines and objects of known sizes (e.g. roads, mountains) provide contextual information that can be used to estimate the scaling that has been applied to an object, thereby affecting our perception of the object itself. As a result, the cars in Figure 6b appear to represent three instances of full-sized cars, despite differences in the size of their representations. Figure 6c shows three cars of the same size. When presented with contextual information suggesting differences in distance, the result is a Ponzo illusion (Figure 6d) where an image of a fixed size appears to have multiple linear sizes based on contextual information. Just as with vowel perception, the different interpretations hinge on differences in the assumed scaling associated with a pattern. Importantly, in both vowel perception and visual object recognition, the disambiguation of some patterns can rely on information external to the differentiating characteristics of the object itself (i.e. 'indirect' effects).

As noted in §1.4, it has been suggested that the many-to-many mapping problem posed by vowel normalization likely involves active control on the part of the listener. If we take into account that vowel perception may involve the consideration of multimodal and contextual 'indirect' information in a speaker-dependent manner, it becomes more difficult to imagine that the process could be controlled in a passive, deterministic manner. The potential importance of considering vowel normalization as a cognitively active process is returned to in the discussion.

1.8. THE CURRENT EXPERIMENT. If vowel normalization operates as a mechanism of perceptual constancy similar to size/distance perception in vision, we expect two general patterns of results related to the estimation of phonemic identity and spectral scaling. First, a sound with a given formant pattern can result in different perceived vowel qualities based on differing estimates of ψ_s. Second, different sounds may have similar vowel qualities based on offsetting differences in apparent ψ_s. These situations are analogous to visual objects that appear to be different sizes despite being represented by the same image (Fig. 6d) and to visual objects that appear to have the same size despite

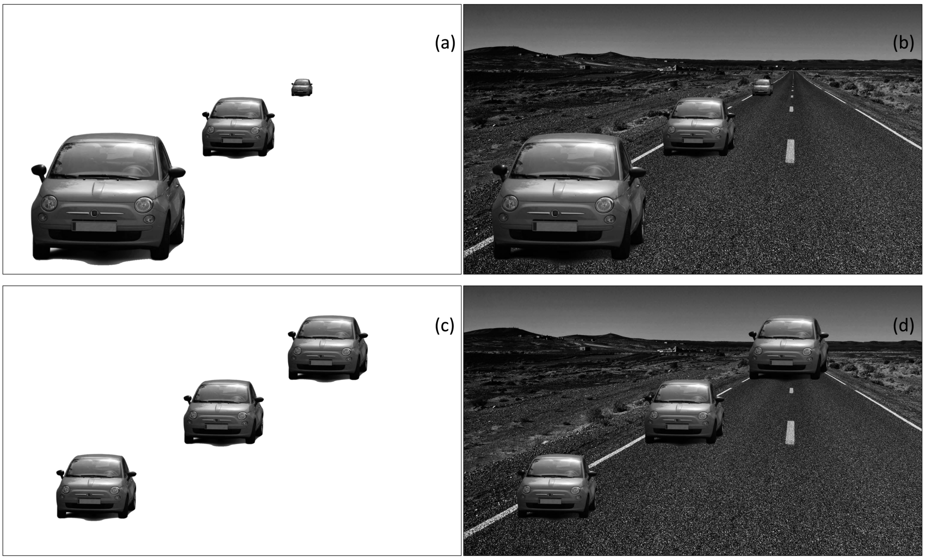

FIGURE 6. (a) Three cars varying according to uniform scaling. (b) When presented against a background suggesting variation in distance the cars from (a) appear to be approximately the same size. (c) Three cars of the same size. (d) When the cars in (c) are presented against the same background as in (b), the result is a Ponzo illusion where a car of a fixed size may appear to be as small as a toy, or as large as a small building, based on external cues regarding distance to the observer.

being represented by different images (Fig. 6b). To investigate this, listeners were presented with a series of vowels varying along a four-step lower-formant (F1, F2) continuum spanning from the /ʌ/ of a low-ψ voice to the /æ/ of a high-ψ voice. Crucially, each successive continuum step features equal proportional changes to F1 and F2. As a result, each step along the continuum is fundamentally ambiguous since it can be interpreted as signaling a difference in vowel quality (V_v), a difference in spectral scaling (ψ_s), or some combination of the two. Each of these steps was mixed with three f0 levels and three higher-formant levels (F3 and above). The effect of f0 on perceived vowel quality was expected to be primarily indirect, by suggesting higher or lower estimates of ψ. Since F3 does not vary in a predictable manner between these two phonemes within-speaker, its effect on perceived vowel quality for these stimuli was also expected to be primarily indirect.

For each trial, listeners were asked to classify the vowel sound as /ʌ/ or /æ/ and to indicate the height of the talker that produced the vowel in feet and inches. Apparent-talker height estimates were recorded in place of ψ estimates because these judgments are highly correlated with ψ (Smith & Patterson 2005) and listeners can provide them in a relatively consistent manner without training. We expected to find two key relationships between apparent-talker height and perceived vowel quality. In cases where acoustically different sounds are associated with a similar perceived vowel quality, we expect offsetting differences in the apparent heights associated with the different sounds, reflecting differences in estimates of ψ. Second, we expect a positive relationship between apparent-talker height and the perceptual 'openness' of the vowel, above and beyond the acoustic characteristics of the sound. This is because when a listener assumes a relatively lower ψ (and therefore a larger speaker), they will also be more likely to identify the vowel sound as /æ/.

2. Methods.

2.1. Participants. Participants were thirty-five native speakers of California English (twenty-eight women and seven men). All listeners reported normal hearing and participated in a single thirty-minute experimental session in exchange for partial course credit.

2.2. Stimuli. Stimuli were thirty-six synthetic vowel sounds, made using a Klatt synthesizer included in Praat (Boersma & Weenink 2016). Each vowel had steady-state formant frequencies and was 225 ms in duration. Stimuli consisted of a four-step lower-formant (F1, F2) continuum, crossed with three higher-formant levels and three f0 levels, for a total of thirty-six stimuli. The low f0 level was set to a value typical for an adult male (120 Hz), while the mid and high f0 levels represented increases of a half octave and an octave, respectively. Final f0 was 20% lower than initial f0 for each level (given in Table 1), and f0 decreased linearly across the duration of the vowel.

Each higher-formant level consisted of a set of fixed, steady-state formant frequencies from F3 to F8. Values for F3 are given in Table 1. The low F3 value was chosen to be appropriate for both /æ/ and /ʌ/ for an average adult male. For the low higher-formant level, every successive formant above F3 had a center frequency 1100 Hz higher than the previous formant. The mid higher-formant level was created by increasing all of the low higher formants by 11%. The high higher-formant level was created by increasing all of the mid higher formants a further 11%, resulting in a 24% difference between the low and high levels.

	LOW	MID	HIGH	
f0	120	170	240	
F3	2475	2755	3068	
STEP	1	2	3	4
F1	684	735	789	848
F2	1354	1455	1563	1679

Table 1. Initial values for f0 and steady-state F3 values for each f0 and higher-formant level, respectively. Steady-state F1 and F2 values for each lower-formant level.

Higher-formant levels were chosen to reflect the natural covariance between f0 and ψ, which suggests that the logarithm of ψ increases at about one third the rate of the logarithm of f0 between speakers (Miller 1989, Nearey & Assmann 2007): Nearey and Assmann report a relationship of $\Delta\log(\psi) = 0.31 * \Delta\log(f0)$ Since f0 levels increased by an octave from low to high, higher-formant frequencies were increased by 24% between the high and low levels since $\exp(0.31 * \log(2)) = 1.24$ The mid higher-formant level was set so that each formant equaled the geometric mean of the each higher formant (e.g. F4, F5, …) at the low and high levels.

2.3. Procedure. Listeners were presented with stimuli randomized along all stimulus dimensions, blocked by repetition. Listeners were played sounds over headphones in a sound-attenuated booth. Each stimulus was presented four times, for a total of 144 trials per subject. Responses were entered using a computer graphical user interface. Listeners were instructed that they would be hearing the voices of males of different ages and were asked to indicate, for each trial: (i) the vowel category, and (ii) the height of the apparent talker. Vowel responses were indicated by clicking on buttons marked ‘HAD’ and ‘HUD’. Height responses were indicated by clicking on a horizontal ruler that spanned from three feet, zero inches to seven feet, zero inches, with markings at three, four, five, and six feet. When listeners clicked on the ruler, the height associated with their response was indicated above the ruler, rounded to the nearest tenth of an inch. Listeners were allowed to replay the stimulus for each trial up to three times. When listeners were satisfied with their responses, they clicked on a button marked ‘submit’ and the next stimulus played after a one-second pause. All listeners completed all 144 trials in the experiment.

The experimental design used here is similar to that used in Barreda & Nearey 2012. However, that experiment focused on the role of f0, and did not find a significant relationship between apparent-speaker size (collected in undefined size units) and perceived vowel quality. The experiment presented here improves on the methods used in that article by collecting apparent height using well-defined units of measurement, using a more powerful analysis method, and minimizing the confound of size and gender by instructing listeners that all speakers were males of different ages.

2.4. Analysis: bayesian multilevel regression models. The experimental design resulted in participants reporting two response variables: vowel quality and talker height. Since one of these dependent variables is dichotomous and the other is continuous, these responses were modeled using two multilevel Bayesian regression models with largely the same structure and predictors. Apparent-talker height for a given trial was modeled as normally distributed with a mean equal to the predicted apparent height for a trial, and a listener-specific variance parameter. Effectively, this means that the residual of height responses is expected to be (approximately) normally distributed. The predicted apparent height for a trial (h) was modeled as varying as a linear combination of several predictors as in equation 4, where β_0^s is the intercept and the remaining β^s terms represent predictor effects for the lower formants (LF), the higher formants (HF), f0, and perceived vowel quality, for subject s. All dependent variables were coded as continuous predictors centered at zero, where differences in levels were indicated by unit changes in the predictors.

(4) $h = \beta_0^s + \beta_{LF}^s\mathrm{LF} + \beta_{HF}^s\mathrm{HF} + \beta_{\mathrm{f0}}^s\mathrm{f0} + \beta_{vowel}^s\mathrm{Vowel}$

(5) $\beta^s \sim N(\mu_\beta, \Sigma_\beta)$

The β coefficients for each listener were modeled as five-dimensional vectors drawn from a multivariate normal distribution (β^s for subject s), as in equation 5. The mean vector, represented by the vector μ_β, represents the equivalent of the ‘fixed effects’, while the listener-specific deviations from these means (β^s) correspond to ‘random effects’ for each predictor. Thus, the models used to analyze the data are equivalent to a mixed-effects model with a maximal random-effects structure, as recommended in Gelman 2005 and Barr et al. 2013.

Vowel responses were modeled as in equation 6, which features the same general model structure outlined above save for two important differences. First, height was used as a predictor in the equation rather than perceived vowel quality (as in equation 4). Second, a logistic regression analysis was carried out so that the model predicted the logit of the probability of observing a ‘HAD’ (/æ/) response on any given trial, as in equation 7.

(6) $\theta = \beta_0^s + \beta_{LF}^s\mathrm{LF} + \beta_{HF}^s\mathrm{HF} + \beta_{\mathrm{f0}}^s\mathrm{f0} + \beta_{height}^s\mathrm{Height}$

(7) $P(\text{response} = \text{‘HAD’}) = \dfrac{1}{1 + e^{-\theta}}$

The model was fit in R (R Core Team 2018) using JAGS (Plummer 2003). Bayesian inference relies on the consideration of the posterior probabilities of model parameters given the data and model structure. The posterior probabilities of parameter values will be summarized using the following statistics: the posterior mean, the posterior standard deviation (analogous to the standard error of the parameter), the two-tailed probability of observing a sample greater/less than zero (analogous to a *p*-value), and the 95% highest-density interval (the smallest interval that encloses 95% of the posterior distribution of a parameter, similar to a confidence interval). Together, these statistics provide information about the most likely values of different parameters, and the uncertainty in these estimates.

3. Results. A summary of apparent-height judgments is presented in Figure 7a, which shows noticeable effects for f0, and only a slight visible effect for the higher formants. Although listeners were allowed to report speaker heights between thirty-six and eighty-four inches, most listeners provided height judgments that fell between approximately fifty-five and seventy-two inches, using only about a third of the possible response range. The restricted range used by listeners can likely be understood in terms of the average heights and average f0 for males of different ages (Figure 7b), which suggests that the low, mid, and high f0 levels resulted in the perception of adult, pubertal, and prepubertal males, respectively.

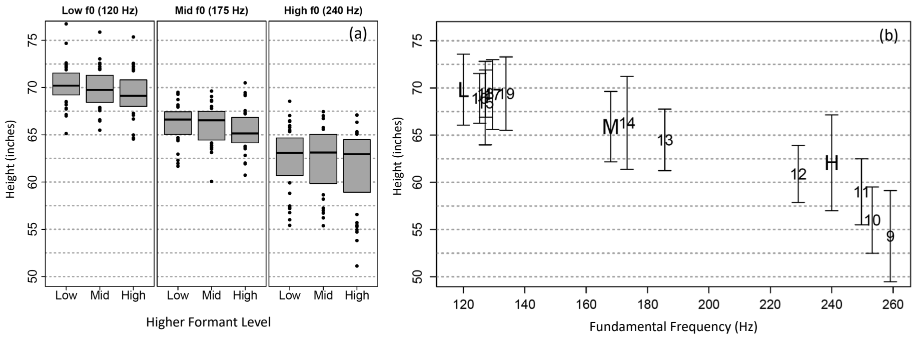

Figure 7. (a) Distribution of average height responses for each listener, averaged across lower-formant level, for each combination of f0 and higher-formant level. (b) Numbers indicate average heights (Fryar, Gu, & Ogden 2012) and f0 (Lee, Potamianos, & Narayanan 1999) of boys of different ages. Error bars indicate one standard deviation. Letters indicate the average height reported for low (L), mid (M), and high (H) f0 levels, plotted according to their initial f0. Error bars indicate one standard deviation of within-listener averages for each condition.

Information about selected model coefficients for the regression analysis of height responses is presented in Table 2. The acoustic effects are all in the expected direction, where increasing formant frequencies and f0s are associated with the perception of shorter speakers. The effect of vowel quality indicates that in situations where the stimulus was classified as /æ/, listeners reported speakers that were approximately 0.9 inches taller on average, holding the acoustic characteristics of the vowel constant. This effect is of a similar magnitude to the spectral effects (−0.71 and −0.38 inches for the lower and higher formants, respectively) but substantially smaller than the effect for f0 (−3.75 inches). The vowel effect is in the expected direction, since a lower assumed ψ is associated with taller speakers and should also result in more /æ/ responses for the lower-formant continuum.

A summary of vowel-quality judgments is presented in Figure 8. Information about regression coefficients for the vowel model is presented in Table 3, where positive coefficients signify an increasing probability of observing an /æ/ response. The LF (lower formant), HF (higher formant), and f0 effects represent changes in log-odds associated with a one-step difference in stimulus level, and all affect the perception of vowel quality in the expected direction. The height effect represents the change in log-odds associ-

effect	mean	*SD*	95% HDI		*p*-value
(intercept)	66.1	0.05	0.07	0.25	0.0013
LF	−0.71	0.06	−0.83	−0.59	< 0.001
HF	−0.38	0.05	−0.48	−0.27	< 0.001
f0	−3.75	0.06	−3.86	−3.64	< 0.001
Vowel	0.90	0.14	0.62	1.17	< 0.001

Table 2. Means, standard deviations (*SD*), 95% highest-density intervals (HDI), and *p*-values for the posterior probability of regression parameters for the speaker-height model. Lower formants (LF) indicate the effects of F1 and F2, while higher formants (HF) indicate the effects of formants F3 and higher.

ated with a one standard-deviation difference in apparent-talker height (5.3 inches across all listeners) and is also in the expected direction: taller apparent-talkers were associated with an increased likelihood of observing an /æ/ response, after controlling for stimulus acoustics.

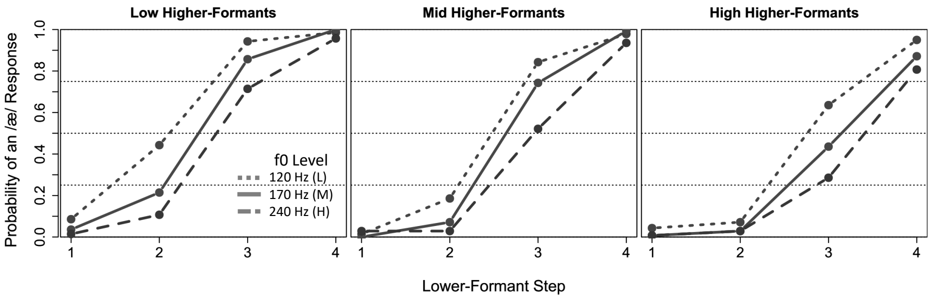

Figure 8. Proportion of /æ/ response for each stimulus across all listeners. Lines within each panel have the same formant pattern and differ only in f0.

effect	mean	*SD*	95% HDI		*p*-value
(intercept)	−1.00	0.09	−1.16	−0.83	< 0.001
LF	4.24	0.16	3.94	4.55	< 0.001
HF	−1.47	0.09	−1.65	−1.29	< 0.001
f0	−0.80	0.11	−1.00	−0.58	< 0.001
Height	0.64	0.11	0.41	0.86	< 0.001

Table 3. Means, standard deviations (*SD*), 95% highest-density intervals (HDI), and *p*-values for the posterior probability of regression parameters for the vowel model. Lower formants (LF) indicate the effects of F1 and F2, while higher formants (HF) indicate the effects of formants F3 and higher.

The statistical models presented in Tables 2 and 3 indicate that apparent height and perceived vowel quality are related after controlling for the formant pattern (F1 through F8) and f0 of vowel sounds. Thus, holding acoustic information constant, in situations where listeners were more likely to identify a sound as /æ/, they were also more likely to hear a taller speaker. Based on the strong correlation between speaker height and vocal-tract length across males of different ages (Fitch & Giedd 1999), and the fact that listeners associate lower ψ with taller speakers, cases where speakers were identified as being taller represent instances where a lower ψ was likely assumed by the listener. Thus, the results presented in Tables 2 and 3 suggest that the spectral scaling inferred by the listener can affect perceived vowel quality, independently of stimulus acoustics.

3.1. Different interpretations of the same sounds. Figure 9a presents the probability of an /æ/ response for all stimulus sounds, divided into the upper and lower quartiles of size responses for each sound. This figure confirms our expectation that the perception of larger speakers will be associated with the perception of more open/

fronted vowels, even when comparing responses for individual stimuli with fixed acoustic properties. Figure 9b presents the average height reported for a subset of stimuli, comparing cases when these sounds were classified as either /æ/ or /ʌ/. We see that listeners reported taller speakers when identifying vowels as /æ/, and that this held true over a range of speaker heights and perceived vowel qualities.

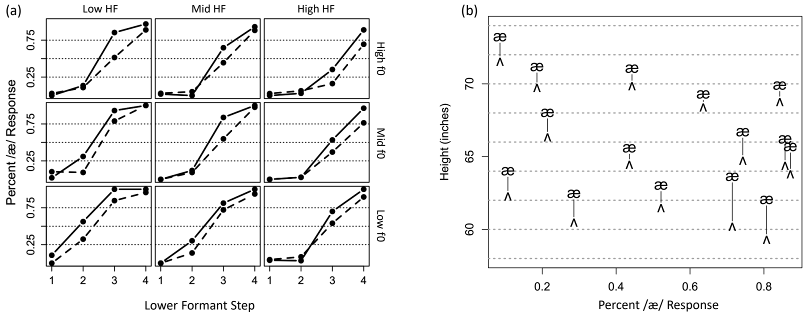

FIGURE 9. (a) Proportion of trials reported as /æ/ across all subjects for each individual stimulus. Lines indicate average responses for upper quartile of size responses (solid line) and lower quartile of size responses (dashed line). (b) Points indicate average height reported for individual tokens when these were classified as either /æ/ or /ʌ/; lines indicate the difference in apparent height for each token according to phonemic classification. Stimuli are arranged on the x-axis based on the rate at which they were classified as /æ/. Only stimuli that were identified as both categories at least ten times each are presented.

One question that arises is whether variation for individual sounds represents differences of opinion between listeners, or whether this variation exists within listener. To investigate this, we considered the cases where listeners were not consistent in their phonemic identifications across their four responses to each stimulus sound. These cases reflect situations where listeners changed their mind about the phonemic identity (and vowel quality) of the sound at least once in the course of the experiment. The difference in apparent height when the listener reported /æ/ compared to when the listener reported /ʌ/ was found for each ambiguous sound, and the average difference between these was found for each listener. This difference (/æ/ − /ʌ/) was 1.15 inches on average across all subjects ($t(34) = 3.3$, $p = 0.002$), an effect of similar magnitude to the 0.9 inch effect of vowel on apparent height reported in Table 2. This indicates that when listeners change their minds about the quality of individual sounds, this is likely to be associated with perceived differences in apparent-speaker height (and ψ).

3.2. SIMILAR INTERPRETATIONS OF DIFFERING SOUNDS. Figure 10 shows the relationship between apparent-speaker height and the probability of an /æ/ response for each stimulus sound, pooled across all subjects. Points joined by lines represent stimuli with the same formant pattern but differing f0s. There are many cases of stimuli with different formant patterns (i.e. points on different lines) falling on nearly the same horizontal position, indicating similar perceived vowel qualities. In most of these cases, the vowel with lower formant frequencies is associated with a taller speaker than the vowel with higher formant frequencies.

The slopes formed by the lines in Figure 10 indicate that differences in f0 led to changes in apparent-talker height, as well as offsetting differences in perceived vowel quality, for any given formant pattern.[6] For example, consider the location indicated by

[6] Deviations from the pattern are mostly for probabilities very near to 0 or 1. The logits of these probabilities can vary dramatically based on very few responses and so are inherently noisier.

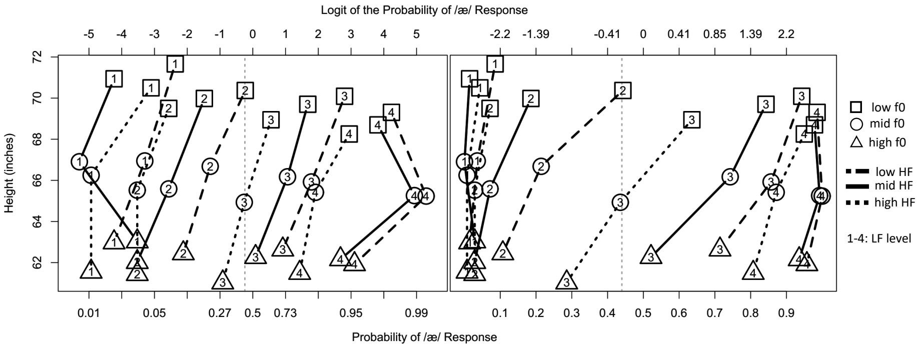

Figure 10. Average height and probability of being identified as /æ/, pooled across all listeners. Each point represents an individual stimulus sound, numbers indicate lower-formant (LF) level, and line type indicates the higher-formant (HF) level. Points joined by lines have the same formant structure but differ in f0. The vertical lines indicate the location of stimuli with a 0.44 probability of being identified as /æ/. On the left, x-axis spacing is in terms of equal differences in logits, while the right panel represents equal differences in probability. The apparent changes in the relationship between apparent height and perceived vowel quality in the right panel arise because of boundary effects at 0 and 1.

the vertical lines in Fig. 10. These lines cross two stimuli that were both judged to be /æ/ approximately 44% of the time, suggesting a similar perceived vowel quality. This is despite the fact that their lower-formant difference (step 2 vs. step 3) resulted in a change of around 50% for most of the categorization functions in Fig. 9a. Crucially, however, in this situation the token with the low higher formants and f0 was identified as being produced by a speaker who was 5’10” tall, while the token produced by the speaker with the mid f0 and high higher formants was identified as being 5’5”. This suggests that vowel quality may be held constant for these acoustically distinct sounds in part due to differing estimates of ψ, which manifest as differences in the apparent height reported for the tokens. In addition, the relationships between apparent height and ψ coincide reasonably well with their empirical values in the population of speakers. The average nineteen-year-old in the United States is 5’10”, while 5’5” represents the average height of children thirteen to fourteen years old (Fryar, Gu, & Ogden 2012). Lee, Potamianos, and Narayanan (1999) report that thirteen- and fourteen-year-old boys produce formant patterns that are approximately 10% higher than those of adult males. This difference in ψ corresponds roughly to the difference between lower-formant steps 2 and 3 (7.5%), the difference offset by the shift in apparent-speaker size between these two tokens.

4. Discussion. The results presented in §3 are consistent with the interpretation of vowel normalization as a mechanism of perceptual constancy where the perception of vowel quality and the perception of apparent-speaker size are related by virtue of a shared reliance on a spectral-scaling parameter (ψ). Some implications of considering vowel normalization in this way are outlined in the remainder of this section.

4.1. Nondeterministic models of vowel perception. Accounts of perceptual vowel normalization can be divided into two general categories based on whether they involve active cognitive control on the part of the listener (Magnuson & Nusbaum 2007). Cognitively active theories feature flexible relationships between inputs and outputs that can be modified as appropriate given the situation, and whose outputs are monitored to improve the performance of the process. In contrast, cognitively passive theories posit static relationships between relevant sensory information and perceived vowel quality. Usually these mappings involve the application of a fixed transformation to a closed set of acoustic inputs so that these can be mapped to a predictable perceived vowel quality. Although cognitively passive approaches to vowel perception may offer useful first approximations to understanding vowel normalization, the lack of flexibility inherent in these approaches does not align with the literature on vowel perception in at least two important ways.

First, it is unclear that there are precise limits on the sorts of information that might influence perceived vowel quality, or speech perception more generally. As noted earlier, nonspectral and even nonacoustic information can affect perceived vowel quality. This behavior is in line with previous reports that speech perception involves the integration of multimodal information such as visual (as in the 'McGurk effect'; McGurk & MacDonald 1976) and tactile information (Gick & Derrick 2009). In the current experiment, we have shown that perceived vowel quality can covary with apparent-speaker height in a predictable manner based on the relationships between vowel quality, speaker size, and spectral scaling As a result, we might expect that a broad range of indirect evidence that suggests speaker size, or correlated indexical characteristics such as age or sex, might affect the perception of vowel quality. However, this is not to say that such indirect information is strictly necessary to the process of vowel normalization. For example, Smith (2014) reports that for vowel sounds of very short durations, listeners are able to categorize vowels accurately in cases where they are not able to reliably identify speaker sex. So, although indirect information appears to influence vowel perception when available, accurate perception can still occur when it is not. As a result, it is difficult to set strict a priori limits on the information that CAN affect speech perception, and on the information that MUST be included in any static mapping between sensory inputs and perceived vowel quality. Although this may seem too flexible, it would be in line with perception in other domains. Holway and Boring (1941) show that observers will rely on different sorts of sensory information to achieve visual size constancy in situations where information is restricted (e.g. in monocular vs. binocular vision). Sperandio and Chouinard (2015) summarize the multimodal nature of perception in the visual domain:

> There is more to sight than meets the eye. Sight can also be the product of multisensory processing … there exists plenty of evidence that our visual perception mediating size constancy is not driven solely by retinal signals but also by extra-retinal signals originating from the eyes and other sensory organs. (Sperandio & Chouinard 2015:278)

Second, it is not clear that information is always used by listeners in a consistent manner regardless of the situation and regardless of speaker expectations. For example, although the fundamental frequency (f0) of vowel sounds affects vowel quality, it does so in a variable manner based on listener expectations (Johnson 1990). Previous experiments have also shown that perceived vowel quality can be affected by presenting listeners with male or female faces (Glidden & Assmann 2004) or simply by instructing listeners as to the sex of the speaker (Johnson, Strand, & D'Imperio 1999). In each case, the nonacoustic information can be thought of as affecting perceived vowel quality indirectly by suggesting more or less plausible estimates of ψ_s for the speaker (as in Fig. 5). Here, results indicate that there can be orderly variation between apparent-speaker size and perceived vowel quality from trial to trial in a situation where listeners are asked to determine their own expectations regarding the speaker, and to report this via an estimate of the speaker's height. In all of these cases, fixed acoustic information relates to different perceived vowel qualities solely as a result of changing listener expectations.

The consequences of these differences can be highlighted by considering a model of vowel normalization that is quite similar to the one presented here, but that nevertheless differs in important ways. Patterson and colleagues (Smith & Patterson 2005, Patterson & Irino 2014) suggest that the peripheral auditory system can automatically estimate S_f, a value analogous to ψ, based on spectral information only, and that this process could underlie vowel normalization and the determination of speaker size:

> the internal version of S_f, functions as an independent variable in the auditory system … . When presented with a pair of speakers, listeners know … which speaker has the longer vocal tract, although they perceive it as a change in speaker size rather than a change in vocal tract length. In short, the experiments support the hypothesis that the auditory system automatically normalizes communication sounds for S_f and [average f0] to segregate the message from the vocal characteristics. (Patterson & Irino 2014:428)

As a result of this peripheral processing, by the time the acoustic information in speech sounds reaches the cortex (where it could interact with linguistic or social information), ψ could already have been estimated and segregated from the representation:

> It is clear that automatic, peripheral normalization of pulse-resonance sounds would greatly benefit communication by adults and the development of communication skills by the young. Central mechanisms of learning, memory and recognition would be insulated from the variability of S_f and [average f0] found in the natural environment, leaving them free to process phonemes in a size-invariant form. This would help explain how children learn speech from a small number of people with widely differing sizes and vocal characteristics. (Patterson & Irino 2014:420)

Reliable, automatically obtainable estimates of ψ would certainly be tremendously beneficial to speech communication. However, it does not seem that this is possible given the between-category overlap that exists in the vowel space. For example, Nearey and Assmann (2007) compared different linear-discriminant models that classified vowels based on the average dialectal formant patterns for each phoneme, and an estimate of ψ for the speaker. Their results indicate that selecting the best-fitting formant pattern leads to worse classification performance, relative to when the plausibility of the values of ψ associated with each classification are also considered. So, listeners appear to rule out potential interpretations if they result in implausible estimates of ψ, even in cases where they would potentially lead to a good fit for a candidate vowel category. This indicates that the spectral information available to human listeners may be fundamentally ambiguous in some cases, such that recourse to 'indirect' information may be necessary for accurate speech perception. However, to the extent that the process of estimating ψ is automatic and cognitively passive, it cannot allow for flexible, context-dependent use of acoustic information to factor into these estimates. It is also difficult to see how nonauditory, indirect information could be integrated into ψ estimates in a cognitively passive manner.

It may once again be instructive to consider perceptual constancy in the visual domain. As noted by Holway and Boring (1941), perceptual constancy represents the limiting case in perception and should not be taken for granted. For example, although infants exhibit visual size constancy (Slater, Mattock, & Brown 1990), adults are better at maintaining size constancy than children are (Zeigler & Leibowitz 1957), indicating that this ability improves with experience. Further, maintaining perceptual constancy is facilitated by knowledge of the characteristics of the objects being observed. The visual system uses the properties of known objects to help improve speed constancy (Martín, Chambeaud, & Barraza 2015), shape constancy (Slater & Morison 1985), color constancy (Granzier & Gegenfurtner 2012), and size constancy (Gogel 1969). In each case, familiarity with the properties of an object helps one to reinterpret apparent changes to the object as environmental or contextual information. In vowel normalization, it is linguistic objects (e.g. phonemes) whose qualities are being held constant by listeners. As a result, linguistic and social/cultural knowledge is likely crucial to the process of vowel normalization. For example, consider the ambiguous location in the formant space presented in Fig. 4a for a language with an /æ/-like vowel but without an /ʌ/-like vowel (e.g. Spanish). Speakers of this language have no reason to produce vowels near /ʌ/, and listeners have no reason to expect vowels near /ʌ/. As a result, the location indicated in Fig. 4a may be less ambiguous in this language as there is only one plausible estimate of phonemic identity and ψ. In this case, linguistic knowledge would help to delimit the interpretations of speaker indexical characteristics.

Initially, we may think of the determination of vowel quality and the communication of social and indexical information as sequential processes such that, first, vowel quality is determined by perceptual processes that may be relatively automatic and 'auditory', and second, normalized vowel-quality units act as the input to the system that determines indexical and social speaker characteristics. So, for example, an instance of /u/ is first recognized as being perceptually fronted, and this information is then used to infer indexical and ideological characteristics associated with participation in the California Vowel Shift. Instead, a consideration of the pattern of results presented in the literature on normalization (and perceptual constancy more generally) suggests that linguistic, contextual, and social knowledge likely play a role in determining perceptual vowel quality. As a result, top-down knowledge is likely involved in the determination of even the most basic units of linguistic structure (cf. Zellou & Pycha 2018). From this perspective, it is not clear that a static mapping that reliably relates acoustics to perceived vowel quality in a manner that is free of context can ever be found. Instead, complete explanations of the determination of perceived vowel quality will likely need to embrace the multimodal and multidimensional nature of speech communication (for an excellent discussion of this topic, see Nusbaum & Magnuson 1997).

4.2. Passive and active listening modes. The reliance on a speaker-dependent parameter in vowel perception suggests that in order to maintain accuracy, listeners' estimates of ψ_s should change when the speaker changes. Magnuson and Nusbaum (2007) suggest that vowel normalization is guided by a process they refer to as 'contextual tuning', which centers around the detection of speaker changes.[7] When a listener encounters a new speaker, they begin normalization procedures to estimate ψ_s for the speaker. This process continues until a stable mapping is achieved, at which point active estimation of ψ_s may stop. The listener may use ψ_s to interpret future speech sounds until they feel that the estimate is no longer appropriate, for example, when a speaker change is detected. When this occurs, the listener may begin to estimate ψ_s for the new speaker. This view of speech perception suggests that listeners may have different listening modes based on the situation (Barreda 2012): a 'new speaker' (active) mode and an 'old speaker' (passive) mode. In the 'new speaker' mode, listeners are estimating ψ_s and carefully monitoring for speaker changes, potentially considering a broad range of information in the process. In contrast, in 'old speaker' mode the listener can passively receive the speech signal with reduced effort since, in the absence of a speaker change, there is no reason to expect ambiguity resulting from sudden changes in ψ.

[7] Nusbaum and Magnuson frame contextual tuning around a speaker-dependent representation, but are not specific regarding the implementation. Thus, the presentation of this theory as centering around ψ, while consistent with their general framework, is adopted to facilitate exposition and is not meant to suggest that Magnuson and Nusbaum necessarily endorse a perceptual framework involving ψ. The integration of contextual tuning with speaker-dependent ψ estimates in vowel perception is presented in more detail in Barreda 2013:119–44, where it is presented as the active sliding template model of vowel perception.

Two kinds of evidence support the notion that listeners behave differently when they are adjusting to different speakers, relative to single-speaker conditions. First, listeners appear to be doing more work, and solving a more difficult problem, when adapting to different speakers. Listeners exhibit longer reaction times (Mullennix, Pisoni, & Martin 1989), lower identification accuracy (Assmann, Nearey, & Hogan 1982), worse serial recall (Martin et al. 1989), and increased neural processing (Wong, Nusbaum, & Small 2004) in situations where they have to adjust to different speakers. Second, listeners can use cues differently depending on whether they need to adapt to different speakers. For example, Nusbaum and Morin (1992) report that listeners rely more on f0 and F3 in mixed-speaker listening conditions, relative to when speech is presented blocked by speaker. Importantly, the different behaviors seen in mixed-speaker listening conditions do not appear to arise from acoustic variation per se, but rather from the realization, conscious or subconscious, that the speaker has changed or may change (Nusbaum & Morin 1992, Magnuson & Nusbaum 2007, Barreda 2012).

After adapting to a speaker, listeners may behave in a manner broadly consistent with a cognitively passive perceptual model featuring static mappings between acoustics and perceptual units. Researchers interested in the transmission of linguistic units in ideal conditions (e.g. two fixed speakers of a single dialect) may be interested primarily in vowel perception in this passive mode. For example, researchers describing phonological or morphological alternations may be primarily interested, and justified, in considering these as static mappings since the effects of between-speaker ambiguity may not be directly relevant at this level of analysis. In contrast, listeners adjusting to new speakers, unknown dialects, or difficult listening conditions may behave in a manner that can only be explained by considering a cognitively active mode of perception. As a result, linguists in other areas, such as those investigating language acquisition or language variation and change, may be interested in considering the consequences of listener behaviors in the active, 'new speaker', mode of perception.

4.3. Implications for the communication and acquisition of vowel sounds. Considering vowel normalization to be an instance of perceptual constancy centering around the uniform scaling of formant patterns offers at least three related benefits for the effective communication and acquisition of speech sounds: it makes it a tractable problem, solvable based on ecologically motivated variation, that allows for solutions based on the joint consideration of a wide range of multimodal information.

First, phone-preserving variation along a single dimension presents listeners with a tractable problem (Nearey & Assmann 2007). When phone-preserving variation in formant patterns is restricted to a single degree of freedom, listeners have strict expectations about what kind of variation is linguistically irrelevant, and what kind of variation is linguistically meaningful. This includes expectations about the locations of vowel phonemes, and expectations about vowel-space dispersion, for any given speaker (Barreda & Nearey 2018). These advantages apply to the uniform scaling of formant patterns along any fixed function of the formant frequencies, whether it be the log scale or an auditory scale such as the ERB scale (Glasberg & Moore 1990). Effectively, the higher the dimensionality of the phone-preserving variation, the more difficult the normalization problem becomes for the listener.

It may be the expectation of variation along a single dimension that allows for the effective communication of vowel quality. For example, suppose that listeners were presented with productions of /u/ from two speakers that differed solely in their F2. Will the listeners think these tokens differ in their perceptual frontness? Under uniform scaling, the independent variation of this formant constitutes direct evidence that one vowel is perceptually more fronted than the other. Thus, if a speaker wants to produce a perceptually more fronted /u/, they understand that independent modification of F2 will convey this information. Conversely, a listener that hears an /u/ with a higher-than-expected F2 can directly interpret this as a difference in the perceptual frontness of the vowel. Similarly, we may consider the perception and communication of hyperarticulation, where a speaker's vowel space is 'larger' than expected. We may ask, larger with respect to what? It is only by defying some expectation regarding vowel-space dispersion that hyperarticulation could result in perceivable differences in vowel quality and communicate social meaning (e.g. D'Onofrio, Pratt, & Van Hofwegen 2019).

We may consider the alternative, that speakers could vary somewhat unpredictably in individual formant frequencies while also producing phonetically identical vowels. This would be as if in the visual domain squares became rectangles as a function of distance in an unpredictable manner, but were somehow still meant to be recognized as squares. If this were the case, the difference in F2 between the two aforementioned tokens of /u/ would be ambiguous: it could represent an intended difference in vowel frontness, or it could simply be an idiosyncratic difference in production that was intended to be linguistically irrelevant and perceptually ignored. In the absence of expectations about the scaling of formant patterns between speakers, a listener would not be able to interpret the difference between these vowels until they had more information about each speaker's vowel space. To our knowledge, it has not been shown that the estimation of unpredictable, idiosyncratic (but still phone-preserving) nonuniform scaling patterns would be a tractable problem for listeners, especially not at the accuracy demonstrated even for syllables presented in isolation. In contrast, Nearey and Assmann (2007) and Turner et al. (2009) have shown that ψ (or an analogous measure) can be accurately estimated from a single vowel even if the vowel category is not known. Further, estimation of a single scaling parameter independently of the inherent 'shape' of an object is an ability that humans (and many other animals) already exhibit in the visual domain when estimating object distance, and in the auditory domain when estimating size information from speech sounds.

Second, uniform scaling as a phone-preserving transformation is not arbitrary, but is instead ecologically motivated. Constraints on phone-preserving transformations of formant patterns may arise from the ecologically determined variation between the size of a resonator and its output formant pattern, and the tendency of humans (and other animals) to interpret this variation as differences in size rather than differences in the content of the message. In fact, this relationship is so robust in the environment that blind listeners also make the association between ψ and speaker height, even making the same sorts of predictable errors as sighted listeners despite never having received any visual confirmation of the relationship between size and formant patterns (Pisanski et al. 2017). In the case of visual patterns, uniform scaling is fixed and determined by the environment. In contrast, as noted in §1.3, the maintenance of uniform scaling by a speech community likely requires effort on the part of speakers and does not occur by default. However, by producing sounds that conform to the ecologically determined covariance pattern between size and formant patterns, vowel normalization can operate via mechanisms already in place to interpret size information from sounds, which likely predate speech communication in humans.

It is important to note that preserving vowel quality in the face of different speaker-specific nonuniform scaling patterns would require a parallel system that is independent of the mechanisms used to infer size from speech. This is because the perception of size from acoustics is understood to be based on differences in uniform scaling of formant patterns (i.e. ψ) due to the relationship between formant patterns and resonator size (Wakita 1977). As a result, it is not clear that the linguistic system can break free of the tendency of listeners to treat uniformly scaled formant patterns as perceptually equivalent. As noted in §1.2, interpreting variation in formant patterns strictly as variation in size presupposes that the objects are recognized as being equivalent, but differing in terms of scaling. Thus, if some nonuniform scaling difference were phone-preserving, listeners would have to simultaneously consider two vowels to be equivalent for the purposes of size perception, but not equivalent for the purposes of vowel quality (or vice versa). Although this is not entirely implausible, theories of nonuniform phone-preserving transformation have a relatively high evidential bar to clear in that they present the listener with several additional problems, go in the face of the ecological variation available to the listener, and provide no clear benefit for speech communication.

Further, in the absence of any external motivation (ecological or otherwise), theories of vowel normalization run the risk of being 'overfit' to a particular set of data. As noted in §1.1, several proposals have been put forth suggesting phone-preserving variation to be quite similar to uniform scaling, but differing somewhat in their details. Researchers propose the best theory given the data at hand; however, data is always limited, leading to slightly different conclusions given the same overall listener behavior. When considering the particular details of a theory of vowel normalization, we may ask, why this way and not another? Although it is not strictly necessary that a proposal have an answer to this question, considering this can help situate an approach within a larger body of evidence and avoid the risk of tailoring the hypothesis to a limited set of observations. Here, we present an approach to understanding vowel normalization with a strong ecological grounding that explains many behaviors seen in speech perception naturally, without recourse to ad hoc mechanisms. Further, this proposal integrates findings from several fields, including human behaviors in other domains and the behavior of other mammals in the auditory domain, most of which are uncontroversial in their own right.

Finally, treating vowel normalization as a problem of perceptual constancy suggests that listeners will naturally use multimodal and 'indirect' information to reduce ambiguity in formant patterns, potentially resolving long-standing issues regarding how listeners identify vowel quality in the absence of context. Phone-preserving variation according to uniform scaling would allow vowel normalization to operate via cognitive scaling operations related to perceptual constancy of the kind that humans are demonstrably adept at resolving, indicating that the cognitive structures or processes required for such tasks are already in place in the brain. Although these benefits would undoubtedly help effective speech communication between adults, they may actually be most helpful in language acquisition since the issue of phone-preserving transformations is particularly acute in this case. For example, how can a child establish a phonemic category when individual tokens can exhibit substantial between-speaker variations in formant patterns? Further, even if the signal the child were exposed to was invariant between adult speakers, it would still be far outside the range of formants that a young child could produce. As a result, the child must translate (i.e. 'normalize') any formant pattern they wish to repeat to a range of frequencies they can produce. Despite these obstacles, six-month old infants recognize that acoustically distinct vowel sounds can nevertheless represent instances of the same phonetic category (Kuhl 1983), and children acquire the speech sounds of their language with relative ease.

Although there is some debate as to what underlies the ability of infants to normalize speech, children might benefit from approaching this as a problem of perceptual constancy. In fact, infants exhibit all of the behaviors necessary to carry out vowel normalization as perceptual constancy. Newborn infants exhibit visual size constancy, demonstrating the ability to carry out the cognitive scaling operations required for vowel normalization (Slater, Mattock, & Brown 1990). Three-month-old infants are sensitive to the relationship between ψ and animal size, indicating that variation according to uniform scaling is understood to represent size information (Pietraszewski et al. 2017). Two-month-old infants are sensitive to mismatches between phonemes and visual evidence of faces producing them (Patterson & Werker 2003), and sex and voice (Bristow et al. 2008), indicating the integration of multimodal information in speech perception. Five-month-old infants use decreasing sound amplitude, along with decreasing image size, to infer increasing distance in visual objects (Pickens 1994), exhibiting the reinterpretation of scaling information as contextual information, rather than indicating a change in the object. In each of these cases, infants are demonstrating that they can solve problems where a given pattern is scaled in a context-dependent manner, and that this variation provides the child with useful information about the perceptual event more generally. These findings suggest that the general cognitive processes required to maintain spectral size constancy (i.e. normalization) either may be innate or are learned early and easily, and could therefore substantially facilitate language acquisition.

5. Conclusion. Our findings show that perceived vowel quality can vary based on differences in the apparent size (and ψ) assumed for a speaker, and that acoustically different vowels can have the same vowel quality when listeners perceive offsetting size (and spectral scaling) differences for speakers. As a result, vowel normalization is broadly similar to size constancy in the visual domain, wherein the ‘shape’ of an object and the scaling applied to it are simultaneously considered in perception. Approaching vowel normalization as a mechanism of perceptual constancy has the potential to help us understand many seemingly disparate listener behaviors such as the use of nonacoustic information, and the role of speaker characteristics in the perception of vowel quality. Further, this proposal is ecologically grounded and fits within a large body of literature from several domains that may be informative to researchers. Here, only a general conceptual framework is presented, with no strong claims about the specific cognitive or neural implementations of the behaviors. Finally, it is important to note that the objects whose properties are being held constant are linguistic in nature and have no independent physical reality. As a result, despite the similarities to constancy in other domains, a full understanding of vowel normalization in perception will necessarily require a deeply ‘linguistic’ accounting of the phenomenon with the consideration of both social and structural aspects of language use.

REFERENCES

Assmann, Peter F., and Terrance M. Nearey. 2008. Identification of frequency-shifted vowels. *The Journal of the Acoustical Society of America* 124.3203–12. DOI: 10.1121/1.2980456.

Assmann, Peter F.; Terrance M. Nearey; and John T. Hogan. 1982. Vowel identification: Orthographic, perceptual, and acoustic aspects. *The Journal of the Acoustical Society of America* 71.975–89. DOI: 10.1121/1.387579.

Barr, Dale J.; Roger Levy; Christoph Scheepers; and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68.255–78. DOI: 10.1016/j.jml.2012.11.001.

Barreda, Santiago. 2012. Vowel normalization and the perception of speaker changes: An exploration of the contextual tuning hypothesis. *The Journal of the Acoustical Society of America* 132.3453–64. DOI: 10.1121/1.4747011.

Barreda, Santiago. 2013. *Cognitively-active speaker normalization based on formant-frequency scaling estimation*. Alberta: University of Alberta dissertation. DOI: 10.7939/R34Q7QZ5N.

Barreda, Santiago. 2017. An investigation of the systematic use of spectral information in the determination of apparent-talker height. *The Journal of the Acoustical Society of America* 141.4781–92. DOI: 10.1121/1.4985192.

Barreda, Santiago, and Peter F. Assmann. 2018. Modeling the perception of children's age from speech acoustics. *The Journal of the Acoustical Society of America* 143: EL361. DOI: 10.1121/1.5037614.

Barreda, Santiago, and Terrance M. Nearey. 2012. The direct and indirect roles of fundamental frequency in vowel perception. *The Journal of the Acoustical Society of America* 131(1).466–77. DOI: 10.1121/1.3662068.

Barreda, Santiago, and Terrance M. Nearey. 2013. Training listeners to report the acoustic correlate of formant-frequency scaling using synthetic voices. *The Journal of the Acoustical Society of America* 133.1065–77. DOI: 10.1121/1.4773858.

Barreda, Santiago, and Terrance M. Nearey. 2018. A regression approach to vowel normalization for missing and unbalanced data. *The Journal of the Acoustical Society of America* 144.500–520. DOI: 10.1121/1.5047742.

Boersma, Paul, and David Weenink. 2016. Praat: Doing phonetics by computer. Online: http://www.praat.org.

Bristow, Davina; Ghislaine Dehaene-Lambertz; Jeremie Mattout; Catherine Soares; Teodora Gliga; Sylvain Baillet; and Jean-François Mangin. 2008. Hearing faces: How the infant brain matches the face it sees with the speech it hears. *Journal of Cognitive Neuroscience* 21.905–21. DOI: 10.1162/jocn.2009.21076.

Broadbent, D. E.; Peter Ladefoged; and W. Lawrence. 1956. Vowel sounds and perceptual constancy. *Nature* 178.815–16. DOI: 10.1038/178815b0.

Charlton, Benjamin D.; William A. H. Ellis; Jacqui Brumm; Karen Nilsson; and W. Tecumseh Fitch. 2012. Female koalas prefer bellows in which lower formants indicate larger males. *Animal Behaviour* 84.1565–71. DOI: 10.1016/j.anbehav.2012.09.034.

Charlton, Benjamin D.; David Reby; and Karen McComb. 2007. Female red deer prefer the roars of larger males. *Biology Letters* 3.382–85. DOI: 10.1098/rsbl.2007.0244.

Charlton, Benjamin D.; David Reby; and Karen McComb. 2008. Effect of combined source (*F*0) and filter (formant) variation on red deer hind responses to male roars. *The Journal of the Acoustical Society of America* 123.2936–43. DOI: 10.1121/1.2896758.

D'Onofrio, Annette; Teresa Pratt; and Janneke Van Hofwegen. 2019. Compression in the California Vowel Shift: Tracking generational sound change in California's Central Valley. *Language Variation and Change* 31.193–217. DOI: 10.1017/S0954394519 000085.

Fant, Gunnar. 1966. A note on vocal tract size factors and non-uniform F-pattern scalings. *Speech Transmission Laboratory—Quarterly Progress and Status Report* 7(4).22–30. Online: http://www.speech.kth.se/prod/publications/files/qpsr/1966/1966_7_4_022 -030.pdf.

Fant, Gunnar. 1975. Non-uniform vowel normalization. *Speech Transmission Laboratory—Quarterly Progress and Status Report* 16(2–3).1–19. Online: http://www.speech .kth.se/prod/publications/files/qpsr/1975/1975_16_2-3_001-019.pdf.

Fitch, W. Tecumseh, and Jay Giedd. 1999. Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America* 106.1511–22. DOI: 10.1121/1.427148.

Fowler, Carol A., and Michael T. Turvey. 1980. Immediate compensation in bite-block speech. *Phonetica* 37.306–26. DOI: 10.1159/000260000.

Fryar, Cheryl D.; Qiuping Gu; and Cynthia L. Ogden. 2012. Anthropometric reference data for children and adults: United States, 2007–2010. *Vital and Health Statistics* 11(252). Hyattsville, MD: US Department of Health and Human Services. Online: https://www.cdc.gov/nchs/data/series/sr_11/sr11_252.pdf.

Gelman, Andrew. 2005. Analysis of variance—why it is more important than ever. *The Annals of Statistics* 33.1–53. DOI: 10.1214/009053604000001048.

Gick, Bryan, and Donald Derrick. 2009. Aero-tactile integration in speech perception. *Nature* 462.502–4. DOI: 10.1038/nature08572.

Glasberg, Brian R., and Brian C. J. Moore. 1990. Derivation of auditory filter shapes from notched-noise data. *Hearing Research* 47.103–38. DOI: 10.1016/0378-5955(90) 90170-T.

Glidden, Catherine M., and Peter F. Assmann. 2004. Effects of visual gender and frequency shifts on vowel category judgments. *Acoustics Research Letters Online* 5.132–38. DOI: 10.1121/1.1764472.

Gogel, Walter C. 1969. The effect of object familiarity on the perception of size and distance. *Quarterly Journal of Experimental Psychology* 21.239–47. DOI: 10.1080/14640746908400218.

Granzier, Jeroen J. M., and Karl R. Gegenfurtner. 2012. Effects of memory colour on colour constancy for unknown coloured objects. *i-Perception* 3.190–215. DOI: 10.1068/i0461.

Hilbert, David. 2005. Color constancy and the complexity of color. *Philosophical Topics* 33.141–58. Online: https://www.jstor.org/stable/43154713.

Hillenbrand, James M., and Michael J. Clark. 2009. The role of f_0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics* 71.1150–66. DOI: 10.3758/APP.71.5.1150.

Hillenbrand, James M.; Laura A. Getty; Michael J. Clark; and Kimberlee Wheeler. 1995. Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America* 97.3099–3111. DOI: 10.1121/1.411872.

Holway, Alfred H., and Edwin G. Boring. 1941. Determinants of apparent visual size with distance variant. *The American Journal of Psychology* 54.21–37. DOI: 10.2307/1417790.

Irino, Toshio, and Roy D. Patterson. 2002. Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform. *Speech Communication* 36.181–203. DOI: 10.1016/S0167-6393(00)00085-6.

Johnson, Keith. 1990. The role of perceived speaker identity in $F0$ normalization of vowels. *The Journal of the Acoustical Society of America* 88.642–54. DOI: 10.1121/1.399767.

Johnson, Keith. 2005. Speaker normalization in speech perception. *The handbook of speech perception*, ed. by David B. Pisoni and Robert E. Remez, 363–89. Oxford: Blackwell.

Johnson, Keith. 2018. Vocal tract length normalization. *UC Berkeley Phonetics and Phonology Lab Annual Report (2018)*, 65–82. Online: https://escholarship.org/uc/item/16c753jz.

Johnson, Keith; Elizabeth A. Strand; and Mariapaola D'Imperio. 1999. Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics* 27.359–84. DOI: 10.1006/jpho.1999.0100.

Joos, Martin. 1948. *Acoustic phonetics*. (*Language* monograph 23.) Baltimore: Linguistic Society of America. DOI: 10.2307/522229.

Kaiser, Peter K., and Robert M. Boynton. 1996. *Human color vision*. 2nd edn. Washington, DC: Optical Society of America.

Kuhl, Patricia K. 1979. Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *The Journal of the Acoustical Society of America* 66.1668–79. DOI: 10.1121/1.383639.

Kuhl, Patricia K. 1983. Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development* 6.263–85. DOI: 10.1016/S0163-6383(83)80036-8.

Ladefoged, Peter, and D. E. Broadbent. 1957. Information conveyed by vowels. *The Journal of the Acoustical Society of America* 29.98–104. DOI: 10.1121/1.1908694.

Lammert, Adam C., and Shrikanth S. Narayanan. 2015. On short-time estimation of vocal tract length from formant frequencies. *PLOS ONE* 10:e0132193. DOI: 10.1371/journal.pone.0132193.

Lee, Sungbok; Alexandros Potamianos; and Shrikanth S. Narayanan. 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America* 105.1455–68. DOI: 10.1121/1.426686.

Lloyd, R. J. 1890. *Some researches into the nature of vowel-sound*. Liverpool: Turner & Dunnett.

Magnuson, James S., and Howard C. Nusbaum. 2007. Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experi-*

mental Psychology: Human Perception and Performance 33.391–409. DOI: 10.1037/0096-1523.33.2.391.

Martín, Andrés; Javier G. Chambeaud; and José F. Barraza. 2015. The effect of object familiarity on the perception of motion. *Journal of Experimental Psychology: Human Perception and Performance* 41.283–88. DOI: 10.1037/xhp0000027.

Martin, Christopher S.; John W. Mullennix; David B. Pisoni; and Walter V. Summers. 1989. Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15.676–84. DOI: 10.1037/0278-7393.15.4.676.

McGurk, Harry, and John MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264.746–48. DOI: 10.1038/264746a0.

Miller, James D. 1989. Auditory-perceptual interpretation of the vowel. *The Journal of the Acoustical Society of America* 85.2114–34. DOI: 10.1121/1.397862.

Mullennix, John W.; David B. Pisoni; and Christopher S. Martin. 1989. Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America* 85.365–78. DOI: 10.1121/1.397688.

Nearey, Terrance M. 1978. *Phonetic feature systems for vowels*. Bloomington: Indiana University Linguistics Club.

Nearey, Terrance M. 1983. Vowel-space normalization procedures and phone-preserving transformations of synthetic vowels. *The Journal of the Acoustical Society of America* 74.S17. DOI: 10.1121/1.2020835.

Nearey, Terrance M. 1989. Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America* 85.2088–2113. DOI: 10.1121/1.397861.

Nearey, Terrance M., and Peter F. Assmann. 2007. Probabilistic 'sliding-template' models for indirect vowel normalization. *Experimental approaches to phonology*, ed. by Maria-Joseph Solé, Patrice Speeter Beddor, and Manjari Ohala, 246–69. Oxford: Oxford University Press.

Nordström, Per-Erik, and Björn Lindblom. 1975. A normalization procedure for vowel formant data. *Proceedings of the 8th International Congress of Phonetic Sciences (ICPhS)*, Leeds.

Nusbaum, Howard C., and James S. Magnuson. 1997. Talker normalization: Phonetic constancy as a cognitive process. *Talker variability in speech processing*, ed. by Keith A. Johnson and John W. Mullennix, 109–32. San Diego: Academic Press.

Nusbaum, Howard C., and Todd M. Morin. 1992. Paying attention to differences among talkers. *Speech perception, speech production and linguistic structure*, ed. by Yoh'ichi Tohkura, Eric Vatikiotis-Bateson, and Yoshinori Sagisaka, 113–34. Tokyo: Ohmsha.

Patterson, Michelle L., and Janet F. Werker. 2003. Two-month-old infants match phonetic information in lips and voice. *Developmental Science* 6.191–96. DOI: 10.1111/1467-7687.00271.

Patterson, Roy D., and Toshio Irino. 2014. Size matters in hearing: How the auditory system normalizes the sounds of speech and music for source size. *Perspectives on auditory research*, ed. by Arthur N. Popper and Richard R. Fay, 417–40. Dordrecht: Springer. DOI: 10.1007/978-1-4614-9102-6_23.

Peterson, Gordon E. 1961. Parameters of vowel quality. *Journal of Speech and Hearing Research* 4.10–29. DOI: 10.1044/jshr.0401.10.

Pickens, Jeffrey. 1994. Perception of auditory-visual distance relations by 5-month-old infants. *Developmental Psychology* 30.537–44. DOI: 10.1037/0012-1649.30.4.537.

Pietraszewski, David; Annie E. Wertz; Gregory A. Bryant; and Wynn Karen. 2017. Three-month-old human infants use vocal cues of body size. *Proceedings of the Royal Society B: Biological Sciences* 284:20170656. DOI: 10.1098/rspb.2017.0656.

Pisanski, Katarzyna; David Feinberg; Anna Oleszkiewicz; and Agnieszka Sorokowska. 2017. Voice cues are used in a similar way by blind and sighted adults when assessing women's body size. *Scientific Reports* 7:10329. DOI: 10.1038/s41598-017-10470-3.

Pisanski, Katarzyna, and Drew Rendall. 2011. The prioritization of voice fundamental frequency or formants in listeners' assessments of speaker size, masculinity, and attractiveness. *The Journal of the Acoustical Society of America* 129.2201–12. DOI: 10.1121/1.3552866.

Plummer, Martyn. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Online: https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf.

Podesva, Robert J.; Annette D'Onofrio; Janneke Van Hofwegen; and Seung Kyung Kim. 2015. Country ideology and the California Vowel Shift. *Language Variation and Change* 27.157–86. DOI: 10.1017/S095439451500006X.

Purcell, David W., and Kevin G. Munhall. 2006. Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *The Journal of the Acoustical Society of America* 120.966–77. DOI: 10.1121/1.2217714.

R Core Team. 2018. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Online: http://www.R-project.org.

Rakerd, Brad, and Robert R. Verbrugge. 1985. Linguistic and acoustic correlates of the perceptual structure found in an individual differences scaling study of vowels. *The Journal of the Acoustical Society of America* 77.296–301. DOI: 10.1121/1.392393.

Reby, David, and Benjamin D. Charlton. 2012. Attention grabbing in red deer sexual calls. *Animal Cognition* 15(2).265–70. DOI: 10.1007/s10071-011-0451-0.

Reby, David, and Karen McComb. 2003a. Anatomical constraints generate honesty: Acoustic cues to age and weight in the roars of red deer stags. *Animal Behaviour* 65.519–30. DOI: 10.1006/anbe.2003.2078.

Reby, David, and Karen McComb. 2003b. Vocal communication and reproduction in deer. *Advances in the Study of Behavior* 33.231–64. DOI: 10.1016/S0065-3454(03)33005-0.

Reby, David; Karen McComb; Bruno Cargnelutti; Chris Darwin; W. Tecumseh Fitch; and Tim Clutton-Brock. 2005. Red deer stags use formants as assessment cues during intrasexual agonistic interactions. *Proceedings of the Royal Society B: Biological Sciences* 272.941–47. DOI: 10.1098/rspb.2004.2954.

Slater, Alan; Anne Mattock; and Elizabeth Brown. 1990. Size constancy at birth: Newborn infants' responses to retinal and real size. *Journal of Experimental Child Psychology* 49.314–22. DOI: 10.1016/0022-0965(90)90061-C.

Slater, Alan, and Victoria Morison. 1985. Shape constancy and slant perception at birth. *Perception* 14.337–44. DOI: 10.1068/p140337.

Smith, David R. R. 2014. Does knowing speaker sex facilitate vowel recognition at short durations? *Acta Psychologica* 148.81–90. DOI: 10.1016/j.actpsy.2014.01.010.

Smith, David R. R., and Roy D. Patterson. 2005. The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *The Journal of the Acoustical Society of America* 118.3177–86. DOI: 10.1121/1.2047107.

Smith, David R. R.; Roy D. Patterson; Richard Turner; Hideki Kawahara; and Toshio Irino. 2005. The processing and perception of size information in speech sounds. *The Journal of the Acoustical Society of America* 117.305–18. DOI: 10.1121/1.1828637.

Smith, David R. R.; Thomas C. Walters; and Roy D. Patterson. 2007. Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled. *The Journal of the Acoustical Society of America* 122.3628–39. DOI: 10.1121/1.2799507.

Sperandio, Irene, and Philippe A. Chouinard. 2015. The mechanisms of size constancy. *Multisensory Research* 28.253–83. DOI: 10.1163/22134808-00002483.

Story, Brad H.; Houri K. Vorperian; Kate Bunton; and Reid B. Durtschi. 2018. An age-dependent vocal tract model for males and females based on anatomic measurements. *The Journal of the Acoustical Society of America* 143.3079–3102. DOI: 10.1121/1.5038264.

Sussman, Harvey M. 1986. A neuronal model of vowel normalization and representation. *Brain and Language* 28.12–23. DOI: 10.1016/0093-934X(86)90087-8.

Syrdal, Ann K., and H. S. Gopal. 1986. A perceptual model of vowel recognition based on the auditory representation of American English vowels. *The Journal of the Acoustical Society of America* 79.1086–1100. DOI: 10.1121/1.393381.

Taylor, A. M.; David Reby; and Karen McComb. 2010. Size communication in domestic dog, *Canis familiaris*, growls. *Animal Behaviour* 79.205–10. DOI: 10.1016/j.anbehav.2009.10.030.

TURNER, RICHARD E.; THOMAS C. WALTERS; JESSICA J. M. MONAGHAN; and ROY D. PATTERSON. 2009. A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data. *The Journal of the Acoustical Society of America* 125.2374–86. DOI: 10.1121/1.3079772.

WAKITA, HISASHI. 1977. Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 25.183–92. DOI: 10.1109/TASSP.1977.1162929.

WINER, ETHAN. 2012. *The audio expert: Everything you need to know about audio*. Boca Raton, FL: CRC Press.

WONG, PATRICK C. M.; HOWARD C. NUSBAUM; and STEVEN L. SMALL. 2004. Neural bases of talker normalization. *Journal of Cognitive Neuroscience* 16.1173–84. DOI: 10.1162/0898929041920522.

ZEIGLER, H. PHILIP, and H. LEIBOWITZ. 1957. Apparent visual size as a function of distance for children and adults. *The American Journal of Psychology* 70.106–9. DOI: 10.2307/1419238.

ZELLOU, GEORGIA, and ANNE PYCHA. 2018. The gradient influence of temporal extent of coarticulation on vowel and speaker perception. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 9:12. DOI: 10.5334/labphon.118.

ZHANG, CAICAI; GANG PENG; and WILLIAM S.-Y. WANG. 2012. Normalizing talker variation in the perception of Cantonese level tones: Impact of speech and nonspeech contexts. Paper presented at Tonal Aspects of Languages—Third International Symposium.

[sbarreda@ucdavis.edu]

[Received 9 April 2019;
revision invited 3 July 2019;
revision received 26 August 2019;
revision invited 2 November 2019;
revision received 27 November 2019;
accepted 15 December 2019]