Journal of Phonetics 63 (2017) 1–18

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Research Article

Listeners respond to phoneme-specific spectral information when assessing speaker size from speech

Santiago Barreda *

University of California, Davis, United States

ARTICLE INFO

Article history:
Received 24 March 2016
Received in revised form 2 January 2017
Accepted 10 March 2017

Keywords:
Vowel perception
Speaker size perception
Speaker characteristics
Speaker normalization

ABSTRACT

Spectral information in speech sounds varies as a function of linguistic content, as well as the vocal-tract length (VTL) of the speaker. It is usually considered that human listeners rely on VTL information when assessing apparent speaker-size. However, a recent experiment (Barreda, 2016) found that listeners respond to the specific spectral-content of speech sounds rather than simply responding to speaker VTL information. This results in biases towards identifying certain phonemes with larger speakers independently of VTL information. To investigate this, listeners were asked to judge relative speaker-size based on vowel pairs differing in vowel quality and/or apparent speaker VTL. Additionally, one group of listeners was asked to report relative-height differences, while another group was trained to report relative-VTL differences directly. Results indicate that both groups of listeners exhibited substantial biases towards associating certain phonemes with larger speakers. In addition, listeners showed substantial variation both in their sensitivity to specific acoustic cues, and in their general approach to speaker size estimation. For example, some listeners rely primarily on VTL cues while others rely heavily on phoneme-specific spectral information.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

In addition to carrying linguistic information, voices carry information that can be used by listeners to infer apparent speaker-characteristics such as speaker gender, age, and size. Although listeners are generally accurate in determining speaker gender from speech (Hillenbrand & Clark, 2009), it has often been noted that they are usually inaccurate in their assessments of speaker size (Collins, 2000; Rendall, Vokey, & Nemeth, 2007; Van Dommelen & Moxness, 1995). In spite of the lack of accuracy, listener judgments of speaker size are usually fairly consistent and predictable on the basis of the acoustic properties of the speech being considered (Bruckert, Liénard, Lacroix, Kreutzer, & Leboucher, 2006; Collins, 2000; Rendall et al., 2007; Van Dommelen & Moxness, 1995). All other things being equal, a token with lower fundamental frequency (f0) will tend to be associated with a larger speaker than a token with higher f0 (Barreda & Nearey, 2012; Rendall et al., 2007; Smith, Patterson, Turner, Kawahara, & Irino, 2005). However, because of its key role in signaling phonemic contrasts, the use of spectral information in the determination of speaker size may be considerably more complicated.

1.1. Vocal-tract length estimates and size-judgments

In general, a speaker with a longer vocal-tract will produce lower formant frequencies (FFs) than another speaker with a shorter vocal-tract (Fant, 1970). Furthermore, vocal-tract length (VTL) is strongly correlated to speaker height across the entire human population, including adults and children of either sex (Fitch & Giedd, 1999). Listeners appear to be sensitive to this pattern of variation and consistently associate lower FFs with larger speakers when linguistic content is controlled across the stimuli being compared (Barreda, 2016; Rendall et al. 2007; Smith et al., 2005). For example, consider a situation where a listener is presented with two instances of /ɑ/ with the same f0 but differing by 15% on average across their FFs. Based on previous experimental results, it is expected that a listener will identify the /ɑ/ with the lower FFs as being produced by the larger speaker. However, as frequently noted (González, 2004; Hollien, Green, & Massey, 1994; Lass &

* Address: 469 Kerr Hall, University of California, One Shields Avenue, Davis, CA 95616, United States.
E-mail address: sbarreda@ucdavis.edu

http://dx.doi.org/10.1016/j.wocn.2017.03.002
0095-4470/© 2017 Elsevier Ltd. All rights reserved.

Brown, 1978; Rendall et al., 2007; Van Dommelen & Moxness, 1995), listeners are not very accurate in identifying the size of adult speakers from speech cues. As outlined in Barreda (2016), this may simply be a result of the fact that when restricted to adult ranges, the amount of systematic variation between size and VTL may be small relative to the amount of variability between speakers. This means that though an underlying systematic relationship between VTL and size may exist in adults given a large enough sample size (Pisanski et al., 2014), this relationship may be overcome by error when any single speaker is considered. However, although listeners are frequently wrong when estimating the size of adult speakers, the consistency of responses observed within and between-listeners highlights a systematic use of spectral information in the assessment of speaker size.

The use of VTL cues in speaker-size judgments is typically investigated by using speech stimuli that contain fixed linguistic content, but vary in apparent VTL. Differences in the apparent VTL of speech sounds are usually simulated by taking speech produced by one speaker (or a small number of speakers) and linearly-scaling the spectral envelope up or down in frequency, resulting in uniform[1] multiplicative increases/decreases in all FFs (Ives, Smith, & Patterson, 2005; Rendall et al., 2007; Smith, Walters, & Patterson, 2007; Smith et al., 2005). Another approach is to use synthetic stimuli, in which case the scaling applied to the formant pattern can be specified directly (Barreda, 2016; Fitch, 1994). Such uniform or nearly-uniform shifts in the spectral content of speech sound are usually thought to affect speaker-size judgments by suggesting differences in speaker VTL, with longer vocal tracts generally implying larger speakers. For example, Rendall et al. (2007) suggest that listeners "discriminate size differences based on formant frequency cues to speaker VTL" (1215). In this view of speaker-size perception, the specific spectral characteristics of a vowel sound, for example as indexed using the FFs, is considered to be informative to speaker-size perception only to the extent that it informs estimates of the speaker's VTL. Although much research on the perception of speaker size relies on listeners having access to accurate speaker-VTL estimates from even short stretches of speech, it is not known if listeners have access to such estimates, or how they might arrive at these.

1.2. Vocal-tract estimation in speech perception

Although many theories of speaker-size perception assume that listeners have access to speaker VTL estimates, VTL information is not directly available in speech sounds and would have to be inferred given the actual formant-pattern present in a sound. However, there are several general theories of speech perception that are compatible with speaker VTL estimation on the part of listeners. Theories of speech perception that assume speaker-dependent interpretation of acoustic information (Barreda, 2013; Joos, 1948; Ladefoged and Broadbent, 1957; Nearey, 1978, 1989), at least implicitly suggest that listeners estimate speaker VTL in the process of speech perception. For example, Joos (1948) suggested that the vowels of different speakers may be "phonetically identical, although acoustically distinct" as long as "each of them occupies the same position within the vowel quadrilateral of the speaker" (p. 59). Although there any many different specific formulations of this general theory of speech perception, what they have in common is that to understand speech the listener must have expectations regarding what range of FFs a speaker is likely to produce. Given that speakers are expected to differ primarily according to VTL within-dialect, committing to a speaker-dependent vowel space with which to interpret vowel sounds is effectively committing to at least a rough speaker-VTL estimate.

For example, consider a vowel sound with an F1 of 600 Hz and an F2 of 1000 Hz appearing on Fig. 1a. This location on the vowel space is closest to /ɑ/ for the long-VTL speaker and /o/ for the short-VTL speaker. Will this vowel be classified as an instance of /o/ or an instance of /ɑ/? If we identify this vowel as /ɑ/, then we must believe that the speaker is large, and if we identify the vowel as /o/, we must believe that the speaker is small. As a result, the vowel quality decision necessarily delimits our VTL (and size) estimate, and vice versa. In this way, theories of speech perception that suggest a speaker-dependent frame of reference necessarily posit a relationship between the identification of speech sounds and VTL estimates for speakers. Based on this relationship, it has been suggested that listeners may recover something like a VTL estimate from the formant-pattern represented in a vowel sound using statistical information regarding the relative locations of vowel phonemes in the dialect (Nearey, 1978; Nearey & Assmann, 2007; Turner, Walters, Monaghan, & Patterson, 2009).

It has also been suggested that speech perception is based on exemplars of previously-experienced speech that are activated in the process of the identification of speech sounds (Goldinger, 1998; Johnson, Strand, & D'Imperio, 1999). According to these theories, details regarding the acoustic characteristics of phonemes are intimately tied to information about the approximate size of the talker that produced them, in addition to other important talker characteristics (age, gender, . . . etc.). Consequently, vowels suggesting roughly the same VTL would be expected to be associated with roughly the same speaker size. For example, under these models the long-VTL vowels in Fig. 1a would tend to be associated with larger speakers (with longer VTLs) by virtue of a lifetime of experience in which the listener has associated low formants with larger speakers. As a result, in practice such an approach to vowel perception makes roughly the same predictions regarding the availability of VTL information in vowel perception as those theories that posit more general speaker-dependent relationships between spectral characteristics and perceived vowel quality.

The above mechanisms would represent cognitive approaches to VTL-estimation that rely on listener knowledge of the sounds of their language, and of the typical characteristics of speech produced by different kinds of speakers. Alternatively, some researchers have suggested that the peripheral auditory system automatically segregates VTL information from phoneme-specific spectral information (Irino & Patterson, 2002; Ives et al., 2005; Patterson & Irino, 2014; Smith & Patterson, 2005; Smith et al., 2005; Turner et al., 2006). In this view, "the auditory system includes an active

[1] For a discussion of the appropriateness of using uniform scaling of formant patterns to simulate differences in VTL between speakers, please see the Appendix of Barreda (2016).

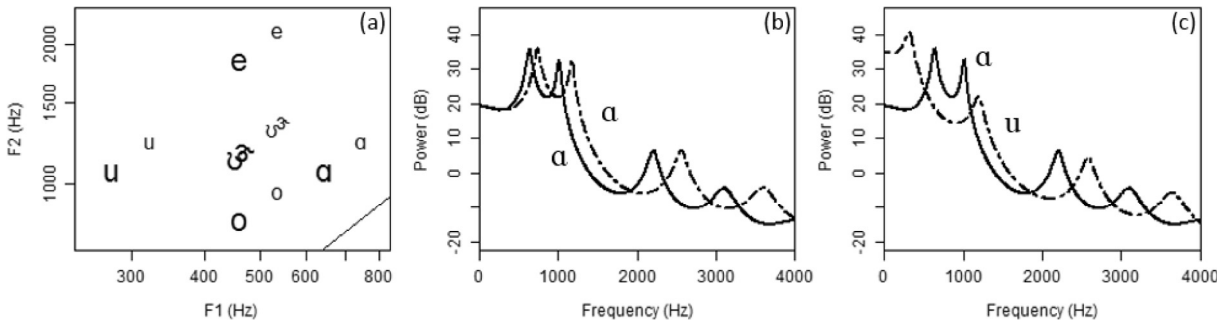

Fig. 1. (a) location of six vowels for a long-VTL voice (large symbols) and a voice with short-VTL (small symbols). (b) Spectra for /ɑ/ for the long VTL (solid line) and short VTL (broken line) vowels. (c) Spectra for /ɑ/ for the long VTL (solid line) and /u/ for the short VTL (broken line).

re-scaling process that is applied to all sounds at an early point in the auditory system [. . .] thereby reducing variability in the [spectral] shape information whilst segregating size information" (Turner et al., 2006, p. 154). Further, this process is hypothesized to form "the basis of the size processing observed in animal behavior." (Irino & Patterson, 2002, p. 305). Under this theory, VTL estimates that are completely free of phoneme-specific spectral information are made automatically available to listeners at a very early stage of processing. If the peripheral auditory system performed such an operation, and listeners had access to the resulting phoneme-independent VTL estimates, then listeners could use the resulting VTL estimates to consistently estimate size for speakers based on VTL information.

The general theories outlined above differ substantially in the manner by which listeners arrive at VTL estimates, and whether the burden is carried by cognitive or physiological systems. Although it is not clear if listeners arrive at VTL estimates using these, or other mechanisms, the important point is that any such process would result in VTL estimates as a by-product of speech perception. As a result, the assumption that listeners estimate VTL in size perception is not unwarranted given some prominent theories of speech perception.

1.3. Size judgments and linguistic content

Although it is not exactly clear if listeners rely on speaker VTL estimates when evaluating speaker size, the utility of such estimates becomes clear when we consider the variability of spectral content across different speech sounds. Fig. 1a presents vowels produced by a speaker with a longer VTL who produces lower FFs overall (large symbols) and the same vowels that have been shifted up by 16% in all FFs to simulate a speaker with a shorter VTL (small symbols). Fig. 1b compares the spectra of the two instances of /ɑ/ on Fig. 1a. If someone is asked to listen to these vowels presented at the same f0 and determine who is taller, they are expected to identify the vowel with the lower FFs and lower spectral energy overall as larger. This expectation also holds for any other same-phoneme comparison that could be made using the voices in Fig. 1a. Results such as these are typically considered to indicate that listeners are responding to the implied VTL differences between the voices.

Research into the determination of speaker-size from speech usually involves designs that control for the linguistic content of speech stimuli, resulting in comparisons as in Fig. 1b. However, in cases where linguistic content is held constant, only VTL differs between voices so that differences in the FF between voices can reasonably be taken as direct evidence of differences in VTL between the speakers. As a result, a listener may simply be comparing the positions of any or all formants across the two sounds and may not be responding to VTL cues at all. Consequently, relative speaker size-judgments from comparisons involving a single phoneme may give the impression that speaker-size judgments are VTL-driven, even in the event that listeners are simply comparing spectral information directly across sounds.

Essentially, experimental designs featuring only same-phoneme comparisons such as those in Fig. 1b do not require that listeners untangle VTL and phoneme-specific spectral information since stimuli feature only VTL variation in spectral content. In order to confirm that speaker-size judgments are driven by VTL rather than direct consideration of spectral information, we would like speaker-size judgments in comparisons featuring different phonemes to also be predictable on the basis of the VTL differences between the vowels, and not on the particular vowels being compared. Importantly, since different-phoneme comparisons actually require that listeners ignore inherent FF differences between vowel phonemes and respond only to VTL differences, different-phoneme comparisons actually provide a strong test regarding the reliance on VTL cues in size perception.

For example, consider possible responses to the vowel pair presented in Fig. 1c. One possibility is that listeners will consistently identify /ɑ/ as taller. They may notice that every single formant (except for F1) is higher for /u/ than /ɑ/, and that since F1 is inherently substantially lower for /u/ relative to /ɑ/, this should not be considered evidence of a longer VTL for /u/. In any case, if a given difference in VTL is implied by the two vowel systems in Fig. 1a, then roughly the same difference in VTL should be implied by the pairs of vowels in Fig. 1b and c. In other words, if the long-VTL voice in Fig. 1a is identified as taller in all same-phoneme comparisons as a result of VTL cues, it seems reasonable to presume that it should also be identified as taller in all different-phoneme comparisons based on the exact same VTL cues.

On the other hand, if listeners simply compare the spectral content of speech sounds when making relative height judgments, they may very well be misled by the inherently-low FFs (and associated low-frequency energy) of certain vowel

sounds. As a result, if listeners estimate relative size directly on the basis of spectral information, then they may very well tend to identify /u/ as taller in Fig. 1c despite the considerably shorter VTL implied by the FFs relative to /ɑ/ in the same figure. This behavior would result in phoneme biases in size perception: the tendency to associate specific speech sounds with larger or smaller speakers independently of the VTL information associated with the sound.

In summary, same-phoneme comparisons feature differences in VTL only and so any association between lower FFs and larger speakers will appear to be VTL driven. Since trials where different phonemes are compared feature phonemically-determined, inherent FF differences in addition to VTL differences, these trials require listeners to ignore phonemic FF variation and only respond to variation associated with VTL differences between speakers. In this way, the absence of phoneme biases in size judgments when comparing different phonemes is a stronger test of the hypothesis that size judgments are driven by VTL rather than simply by specific spectral information.

1.4. Phoneme biases in size perception

Barreda (2016) presented listeners with pairs of synthetic vowel sounds that varied in apparent VTL and/or vowel quality. Listeners were asked to indicate which of the vowels sounded like it had been produced by the taller speaker. This approach is similar to that employed in previous experiments investigating the perception of speaker size (Rendall et al., 2007; Smith et al., 2005). However, unlike in previous experiments, linguistic content was not controlled for and the vowels in the pair could be any of /æ/, /i/ or /ʊ/. These vowels feature phonemically-determined, inherent variation in their FFs that is large enough to potentially overwhelm the more subtle differences in the FFs typically associated with variation in VTL. For example, /ʊ/ had F1 and F2 frequencies that are 64% and 43% lower than those of /æ/, even at a single VTL level. In contrast, the largest VTL difference in the experiment was associated with a 16% difference in all FFs (approximately the average difference between adult males and adult females).

In same-phoneme trials, listeners were asked to compare the same phoneme at two VTL levels. As in previous experiments featuring these sorts of comparisons, responses were largely predictable on the basis of the VTL differences between the voices. However, in trials where listeners were asked to compare instances of different phonemes (different-phoneme trials), listeners responded to both the VTL differences between the vowels, and to the inherent spectral characteristics of the vowels being compared. For example, /ʊ/ was consistently identified as taller than /æ/, even when presented at a shorter apparent VTL.

The existence of a tendency towards associating towards associating some vowel phonemes with smaller or larger sizes even when in conflict with VTL information is problematic for any purely VTL-driven approach to understanding speaker-size judgments, regardless of how these estimates are made available to listeners. The presence of phoneme biases suggests that speaker-size judgments may involve a direct consideration of the spectral energy in speech sounds. Furthermore, the results presented in Barreda (2016) indicate that phoneme biases may be so large as to overwhelm the VTL information in a speech sound, thereby dominating the perception of speaker size from spectral information. The presence of phoneme biases in size judgments would also significantly alter the manner in which size perception is investigated. As mentioned earlier, these investigations typically control for linguistic content, or only consider aggregate judgments of speaker size across a fixed set of stimuli for all speakers. Either of these approaches may obscure the process of size perception by giving the impression that listeners are responding to VTL cues when they are in fact simply responding to the specific characteristics of the stimuli being presented. For example, had the results of Barreda (2016) only been modelled on the basis on VTL cues, the existence of phoneme biases in size judgments would not, and could not, have been found. It is only by investigating how size judgments vary as a function of the specific spectral content in a vowel sound (i.e., the inherently-low FFs of /ʊ/, independent of VTL) that we may observe how size perception varies as a function of specific spectral information. Consequently, the presence of phoneme-biases suggests that size perception should be investigated by considering size judgments on the basis of the actual spectral characteristics of the stimuli being presented, and not just using VTL information.

1.5. Rationale for the current experiment

The results presented in Barreda (2016) suggest that listeners do not base their speaker-size estimates solely on VTL estimates, and that phoneme-biases may play an important role in speaker-size perception. However, the limited number of phonemes used (/i æ ʊ/) did not allow for an investigation into the nature of these biases, how they are influenced by the particular formant-patterns of the vowels being compared, or how sensitive listeners are to inherent FF differences when making size judgments. The experiment described here used a larger stimulus set consisting of six vowel phonemes (/ɑ e o ɝ u ʊ/). Because of their respective positions in the vowel space (Fig. 1a), these vowels can be arranged in pairs such that different-phoneme pairs feature a difference in VTL in addition to differing in only one or two individual formants at a time. Essentially, the idea is to create situations where listeners are expected to exhibit phoneme biases by asking them to compare vowel phonemes with large inherent (phonemically-determined) differences in their FFs. Furthermore, by comparing the difference in relative-size judgments between situations where vowels differ only in VTL (same-phoneme trials) to situations where vowels differ in VTL and a limited number of formants (different-phoneme trials), this design allows for an investigation into how inherent differences in FFs across phonemes result in phoneme biases in a relatively controlled manner.

Furthermore, this experiment seeks to investigate the use of spectral information in speaker-size judgments by comparing two groups of listeners: one group that reported differences in speaker height, and another that was trained to report VTL differences (i.e., uniform shifts in FFs) between speakers directly. The phoneme-biases presented in Barreda (2016) raise questions regarding whether listeners estimate speaker VTL in the process of making speaker size judgments, as is

at least implicitly assumed in most research regarding the perception of speaker size. However, it may be the case that listeners do have access to VTL estimates for speakers, but that relative speaker-size judgments involve the joint consideration of these estimates and additional spectral information, leaving them susceptible to phoneme biases. In order to tease apart these possibilities, it is necessary to collect VTL judgments from listeners directly.

Unfortunately, in different-phoneme trials featuring significant phoneme-biases, size judgments may no longer act as effective substitutes for VTL-estimates. In these situations, it would be useful if VTL judgments could simply be provided directly. However, unlike the perceptual quality associated with differences in f0 (pitch), there is no commonly-accepted term in the English language for the perceptual quality associated with VTL differences. For example, helium speech typically makes any given speaker sound smaller and untrained listeners may refer to this speech as having a higher than normal pitch. However, helium speech is not characterized by a higher than normal f0. Instead, helium raises the resonance frequencies of the vocal tract by increasing the speed of sound, resulting in roughly uniform scaling of the spectral envelope of the kind typically associated with decreases in the VTL of a speaker (Podhorski, 1998). Although listeners may be generally aware of the perceptual quality associated with VTL differences and have been demonstrated to use it to make absolute and relative size judgments, the lack of a commonly-accepted term for this characteristic makes it difficult to collect VTL estimates in the absence of an appropriate proxy judgment (e.g., size). To this end, a group of listeners was trained to provide VTL judgments directly using the method outlined in Barreda and Nearey (2013a).

The necessity of training listeners to report VTL, and the amount of training listeners should require, will depend on the availability of VTL estimates to listeners. If VTL estimates are provided automatically to listeners by some physiological or cognitive process, and these estimates form the basis of size perception, then listeners should not require much training at all. If this were the case, learning to report VTL directly would simply be a case of learning a label for a value that listeners have direct access to and which already guides their estimation of speaker size. On the other hand, if VTL is not easy for listeners to estimate or report, they may require extensive training before they could be expected to accurately report VTL. Although this may seem to be a weakness of a reliance on trained-listener VTL judgments, note that this outcome would be extremely problematic for the general view that size judgments are driven by VTL estimates. If substantial training were necessary before people could report VTL reliably, and these same VTL estimates were the foundation of size judgments made by human listeners, then it would follow that human listeners should also require extensive training before they could be expected to report size reliably. Adopting this position would be putting the cart before the horse however, as it would suggest that size judgments in nature are 'wrong' and must be corrected in a lab before they can be accurately collected. Instead, we must consider that if listeners require extensive training before they can report VTL accurately and consistently then perhaps listeners do not have ready access to these estimates, and do not base their judgments of speaker size on them.

The experiment outlined below will compare responses from two groups of listeners, one group providing 'naive' relative-height judgments, and one group trained to provide relative-VTL judgments directly. By comparing judgments across these two groups we may investigate the role of VTL estimates in relative speaker-size judgments in more detail. We may consider three general possibilities in comparing the results of the trained and untrained group of listeners. First, listeners in both groups may simply respond to VTL differences between voices and not exhibit strong biases towards identifying some phonemes as larger/taller independently of VTL information. These results would run counter to the persistent phoneme-biases reported in Barreda (2016). Second, listeners in the trained group might respond primarily to VTL differences, while those in the untrained group might exhibit large phoneme biases. This result would indicate that listeners can recover phoneme-independent VTL estimates, but that perhaps the size judgments made by listeners involve the consideration of additional spectral information leaving them susceptible to phoneme biases. Finally, both groups of listeners might show large phoneme biases in their responses (relative height or relative VTL). This result would suggest that listeners do not have ready access to VTL estimates in a manner that they may be reported accurately and consistently, independently of linguistic content.

2. Materials and methods

2.1. Participants

Participants were 46 students from the University of Alberta drawn from a participant pool in which undergraduate students take part in experiments in exchange for partial course credit. All participants were students taking an introductory level, undergraduate linguistics course, and were native speakers of English. Listeners were randomly assigned to one of two groups: a training group (23 listeners), and a control group (23 listeners). The 23 listeners who were trained to report VTL performed 12 minutes of the training method described in Barreda and Nearey (2013a,b).

2.2. Stimuli

Stimuli consisted of synthetic vowels based on average productions of adult male speakers of Edmonton English. The first four FFs for the longest-VTL voice are presented in Table 1. F4 and the higher formants were fixed across vowel categories at appropriate values for an adult male speaker, following conventions when using synthetic vowels (Klatt, 1980; Nearey, 1989). Each consecutive formant above F5 was set to 925 Hz higher than the previous one, up to the 11th formant to prevent differences in spectral slope associated with varying distances between the highest specified formant and the Nyquist frequency (Holmes, 1983). Formant bandwidths were fixed at 6% of formant center frequencies, with a minimum bandwidth of 60 Hz. All vowels had steady-state formant frequencies and were 200 ms in duration. The fundamental frequency for each vowel decreased linearly from 120 Hz to 110 Hz from the beginning to the end of the vowel. Vowels were synthesized using a Klatt-style (Klatt, 1980) parametric

Table 1
Lowest four formant frequencies for stimulus vowels for the longest-VTL voice.

Vowel	ɑ	e	o	ɝ	ʊ	u
F1	646	462	462	462	462	277
F2	1062	1846	831	1154	1154	1062
F3	2308	2308	2308	1662	2308	2308
F4	3139	3139	3139	3139	3139	3139

synthesis program implemented in MATLAB. After synthesis and prior to experimentation, speakers of the local dialect confirmed that the resulting stimuli were appropriate exemplars of their respective phonemic categories.

The vowels at a given VTL level can be thought of as a plausible set of vowels produced by a single synthetic 'speaker' with a single VTL, and complete with stable higher-formants within-speaker. Simulated VTL differences were created by increasing the FFs of the longest-VTL vowel stimuli (Table 1) uniformly in logarithmic steps, and synthesizing the resulting vowels as described above. The FFs of the longest-VTL voice were log-transformed, increased by 0.04, 0.12 and 0.16 log-Hz to all FFs, and then exponentiated. These log-Hz increases correspond to changes of 4.1%, 12.7% and 17.3% relative to the values presented in Table 1. The result of this was that the voices with the shortest and longest-VTL were separated by about 17% in their FFs (0.16 log-Hz), while the middle two voices were separated by about 8% in their FFs (0.08 log-Hz). A difference of about 16% in all FFs is approximately the difference between adult males and adult females, while also being roughly the magnitude of within-category variation for adult males exhibited in large vowel data sets (Hillenbrand, Getty, Clark, & Wheeler, 1995; Peterson & Barney, 1952).

The six stimulus vowels for the longest and shortest-VTL voices are presented in Fig. 2a. Each of the different-phoneme pairs used in this experiment highlights a difference in one or two FFs in addition to VTL differences: /u/-/ɑ/ was used to investigate the effect of F1, /e/-/o/ was used for F2, /ɝ/-/ʊ/ was used for F3, /o/-/ɑ/ was used for F1 and F2 together, and /ɝ/-/e/ was used for F2 and F3 together. Examples of the five different-phoneme pairs used in this experiment are presented at the large VTL difference (0.16 log-Hz) in Fig. 2b–f. For each pair, the vowel with inherently-lower FFs is shown at the short-VTL level. This highlights that phonemically-determined FF differences between vowels can easily overwhelm the relatively subtler FF differences associated with VTL differences between speakers.

2.3. Training

The VTL training method used for this experiment is outlined in detail in Barreda and Nearey (2013a); a summary will be provided here. The computer-based training program consisted of a rectangular board (900 by 700 pixels) presented on a computer monitor, where different locations on the board were associated with different voices. Voices were arranged on the board in a grid so that they differed horizontally in apparent VTL (implemented by scaling the FFs up or down), while voices differed vertically in terms of f0. All voices were represented by the vowels /i/ and /æ/, played together after a brief silence. There were 100 rows, with every row having an f0 level 1.1% higher than the level below it, and 40 columns with every column having a formant pattern that was 1.2% higher than the one to its left. In this way, the voices associated with different locations on the board spanned roughly the entire range of f0 and VTL seen in the human population, with f0 spanning from 100 Hz to 300 Hz, and FFs ranging from very low values (e.g., the first three FFs for /i/ being 275, 2114 and 2711 Hz respectively) to values 58% higher.

The high number of individual voices (4000) and their relative proximity both in acoustic terms and on the board (adjacent voices were separated by 20 pixels horizontally and 6 pixels vertically) were meant to give the listener the impression of a continuous response space. For the sake of comparison, just-noticeable differences for VTL cues (i.e., uniform scaling of the formant pattern) have been estimated to be 7–8% for isolated vowels (Smith et al., 2005) and 4–6% (Ives et al., 2005) for syllable phrases. In both Smith et al. (2005) and Ives et al. (2005), just noticeable differences were estimated using a two-alternative forced-choice methodology. No markings were made on the board to indicate the exact spaces associated with any given voice. The layout of voices on the board meant that the bottom-left corner represented voices roughly appropriate for adult males (low f0–low FFs). As one proceeds diagonally towards the top-right corner of the board, f0 and the FFs increase in unison so that eventually voices are appropriate for adult females (mid f0–mid FFs) and then young children (high f0–high FFs). However, because of the crossed stimulus design, the voices spanned a wide range of combinations including more common ones (low f0–low FFs, high f0–high FFs), but also less common ones towards the top left (high f0–low FFs) and bottom right (low f0–high FFs) corners of the board.

The training procedure was as follows. Listeners were told that each part of the board was associated with a different kind of voice. They were told that voices differed in pitch from bottom to top, and in 'voice size' (i.e., VTL) from left to right. Listeners were instructed that 'voice size' was something there was no term for in English, but that it was an acoustic characteristic that was independent of pitch. First, listeners were allowed to familiarize themselves with the voices on the board by clicking on the board 25 times, where the nearest voice on the board played after every click. Listeners were asked to use this time to note what 'voice size' sounds like, for example by moving horizontally at a fixed height (f0/pitch level) when they clicked.

After this, listeners moved onto the training portion, which they performed for 12 min. Listeners heard a randomly-selected voice from the 4000 stimulus voices and were asked to try to identify its location on the board using two guesses. In identifying the location of the voice on the board, listeners were effectively providing VTL and f0 estimates for the voice. After the listener provided their first guess, the location of this guess

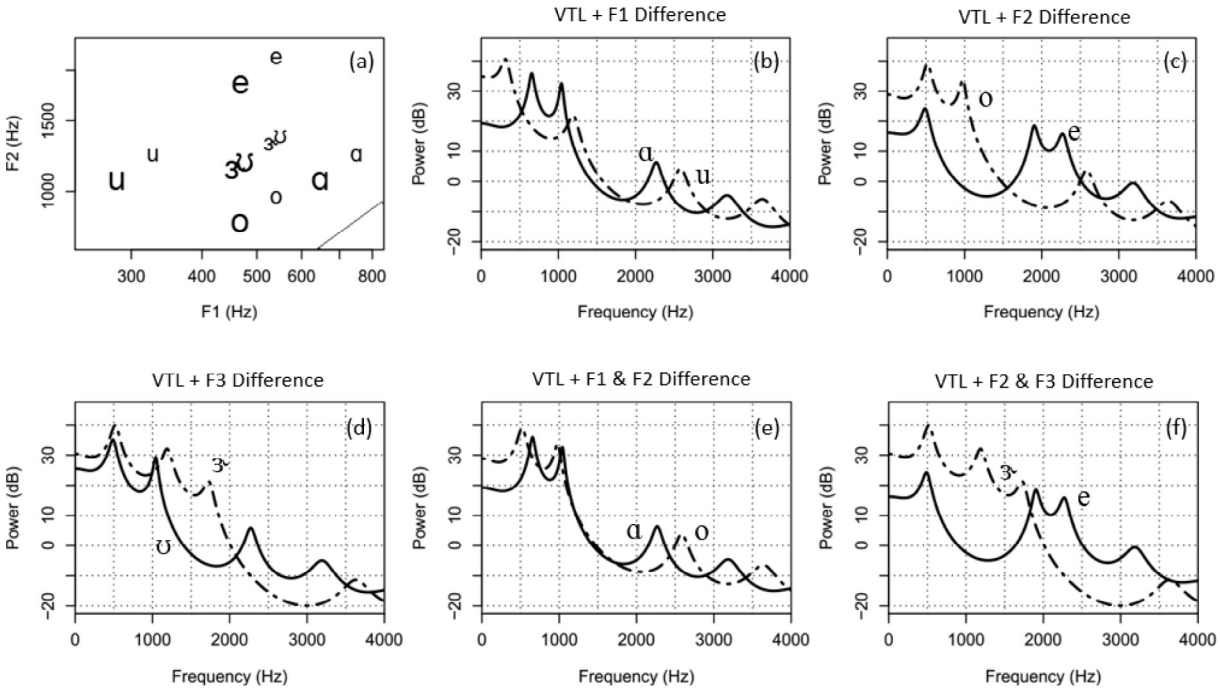

Fig. 2. (a) The six stimulus vowels for the longest (large symbols) and shortest VTL levels (small symbols) presented on a formant space. /ɜ/ and /ʊ/ have been adjusted slightly so that they do not entirely overlap. (b–f) Examples of the different-phoneme pairs used in this experiment. In all cases, the dash-dot line indicates the vowel with the shorter simulated VTL.

was indicated with a point on the board, and the nearest voice on the board was played for them. This was done to allow listeners to refine their guess using their memory of the stimulus voice and the characteristics of the location of their guess. They were then allowed to provide a second guess, which was also indicated on the board. After the second guess was provided, a bulls-eye indicated the actual location of the stimulus voice, which allowed the listener to compare the location of their guesses to the actual location of the voice on the board. In this way, the training method provided both visual and auditory feedback to listeners to help them understand the organization of the space and to help them improve their ability to report apparent VTL directly. The correct location of the voice on the board was displayed for one second, after which the board was cleared and the next voice was played. However, the training method was self-paced in that listeners were not given any time limits regarding how quickly they should be providing their first or second guesses for each voice.

2.4. Procedure

Listeners were told that they would be hearing a series of synthetic voices patterned after adult-male speakers. Listeners were instructed that they would be hearing the vowels in the words ‘who’, ‘haw’, ‘hood’, ‘heard’, ‘hoe’, ‘hey’ presented in pairs. Listeners in the control group were asked to report the relative difference in speaker height, while listeners in the training condition were asked to report the difference in ‘voice size’ (i.e., VTL) between the speakers. Unless a distinction specifically needs to be drawn between these responses, these relative height and VTL judgments will collectively be referred to as relative-size judgments.

Sounds were presented over headphones, in a sound-attenuated booth. Vowel sounds were presented to listeners in pairs separated by 300 ms of silence. Listeners were asked to make a relative-size judgment regarding the speakers who produced the vowel sounds. Listeners provided responses by clicking on a specially-designed graphical user interface that contained a slider that allowed participants to report relative-size differences on a continuous scale. Listeners were told that moving the slider to the left would indicate that the first voice was bigger, while moving the slider to the right would indicate that the second voice was bigger. Further, listeners were instructed that the extent to which the slider was moved towards one extreme or the other should reflect the magnitude of the relative size difference between the voices. For example, a slight deviation off-center to the left would indicate that the first voice was judged to be somewhat larger than the second, while a larger deviation should reflect a larger perceived difference. The slider was 700 pixels in length, and the final position of the slider was recorded as the response. This value was centered and flipped in sign so that 0 meant equal-sized voices, positive values meant the first voice was larger, and negative values meant the second voice was larger. Two labels appeared over the slider, one over the left half and one over the right half. These labels read 'HOOD', 'HEARD', 'HEY', 'HOE', 'WHO', or 'HAW' as appropriate given the vowels in the trial, and were simply intended to remind the listener of the order of the sounds in the pair. The user interface contained a button marked 'Replay' that allowed listeners to replay the presented stimuli up to three additional times for each trial. After the listener had made a selection, they had to press a button marked ‘Submit’, and the next stimulus was played after a one second pause.

In both same- and different-phoneme trials, the first stimulus VTL level was always presented with the fourth (a 0.16 log-Hz difference), and the second VTL level was always presented against the third (a 0.08 log-Hz VTL difference). The six vowel

categories were presented with themselves at the two VTL difference levels, and balanced for order with respect to the VTL difference between the voices. This resulted in 24 unique same-phoneme vowel pairs. Each of the 5 different-phoneme pairs was also presented at each of the two VTL difference levels, and balanced for order with respect to VTL difference and vowel category, resulting in 40 unique different-phoneme vowel pairs. These 64 total vowel pairs were presented to listeners randomized along all stimulus dimensions, and blocked by repetition. Listeners completed a maximum of 6 repetitions (n = 384), with a self-timed pause between the third and fourth blocks. Listeners participated for a maximum of one hour, meaning some listeners did not complete all six repetitions.

2.5. Statistical analysis: Bayesian multilevel linear regression

The experimental task asked listeners to report relative-size differences between the vowels in each stimulus pair resulting in a continuous dependent variable. Results were analyzed using a Bayesian multilevel linear-regression model. This approach simultaneously models the results of individual subjects, while also pooling information across all subjects to estimate group-level effects. Because of the relatively large number of observations per listener, an alternative approach may have been to fit a model for each listener and then carry out significance testing using the distribution of estimated parameters across subjects (Gumpertz & Pantula, 1989; Lorch & Myers, 1990). Although this 'no-pooling' approach (Gelman & Hill, 2006) is simpler, estimating parameters within a multi-level model offers several advantages. For example, the pooling of information across subjects results in shrinkage which pulls parameter estimates towards group means, thereby offering protection against false positives and allowing for multiple comparisons without the resizing of confidence intervals for different parameters (Gelman, Hill, & Yajima, 2012; Kruschke, 2010, 2014; Kruschke, Aguinis, & Joo, 2012).

Bayesian inference relies on consideration of the posterior distribution of parameter values given the data and the prior probabilities of the parameters. These distributions may be approximated using Markov Chain Monte Carlo methods, which find a distribution of jointly-credible values for all model parameters by taking a series of random 'steps' through the joint parameter space. The result of this is a 'chain' of parameter values which can be used to assess credible, and jointly-credible, values for parameters or combinations of parameters. As a result, Bayesian estimation provides confidence intervals for all estimated parameters so that the values of different parameters (or combinations of parameters) may be easily contrasted. If a value of interest (e.g. 0) is typical given the distribution of values in the posterior distribution (the 'chain') it is deemed to be a credible value for that parameter. On the other hand, if the value of interest is not typical given the values in the chain, it is deemed to not be a credible value for the parameter (for more information on Bayesian model-fitting and inference please see: Gelman & Hill, 2006; Kruschke, 2014; Kruschke & Vanpaemel, 2015).

2.5.1. Model and estimation details

Modelling relative-size judgments using specific stimulus characteristics poses a challenge in that absolute FFs will necessarily be highly correlated with any spectrally-based measure of VTL. As a result, absolute FFs and VTL information cannot both be used as predictors in the model. The primary research question involves the investigation of phoneme-biases: effects for phoneme-specific information on size judgments above and beyond VTL cues. As a result, VTL cues were given a priority within the model and formant information was only specified when these deviated from what would be expected given the VTL cues.

All vowel pairs featured a difference in scaling of either 0.08 or 0.16 log-Hz. This means that if the same phoneme were being presented at two different VTL levels, any given formant would differ between the two vowels by either 0.08 or 0.16 log-Hz (approximately 8.3% and 17.3%). The effect for VTL differences between voices was included in the model using the difference in scaling between the voices expressed in log-Hz, relative to the first voice in the pair. Negative values of this predictor ($\Delta VTL = VTL_1 - VTL_2$) indicate that the first voice had a longer VTL and lower FFs overall. Differences between the first three formants were calculated relative to their values at a single VTL level (e.g., the values in Table 1). These differences were expressed using the difference in log-Hz between the two formants relative to the first vowel in the pair. For example, F1 was 6.47 log-Hz for /ɑ/ (646 Hz) and 5.62 log-Hz for /u/ (277 Hz), meaning that the predictor associated with differences in F1 across the vowels (ΔF1) would equal 0.85 when /ɑ/ was the first vowel in the pair, and -0.85 when it was second. This design means that values of the formant difference predictors (ΔF1, ΔF2, ΔF3) were all equal to zero in same-phoneme trials, and only took on non-zero values in different-phoneme trials. The model also includes formant-difference interactions (ΔF1F2, ΔF2F3), entered as the cross-product of their constituent formant-difference terms.

The result of this coding scheme is that the VTL predictor (ΔVTL) represents coordinated, uniform shifts in all FFs of the kind associated with VTL differences between speakers. In other words, the ΔVTL predictor reflects the kind of spectral information that is usually thought to guide speaker-size judgments. On the other hand, the individual FF difference predictors and their interactions represent inherent, between-phoneme, within-speaker variation in formant patterns and do not in any way reflect the VTL difference between the vowels in a pair.

Prior to analysis, relative-size judgments were scaled within-subject so that the standard deviation of responses for each subject was equal to 1. Responses were modeled as coming from a normal distribution with an unknown mean, and a listener-specific error term (for j listeners), as in Eq. (1). Listener-specific error terms were used in order to accommodate the differences in the systematicity of responses between the different listeners. Each of these error terms was given a uniform prior

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

$$\mu = \alpha_0 + L + \beta_{\Delta F1}\Delta F1 + \beta_{\Delta F2}\Delta F2 + \beta_{\Delta F3}\Delta F3 + \beta_{\Delta F1F2}\Delta F1F2 + \beta_{\Delta F2F3}\Delta F2F3 + \beta_{\Delta VTLS}\Delta VTL + \beta_{VTLD}\Delta VTL \times \text{Different} \tag{2}$$

The relative-size difference between the vowels on a given trial (μ) was modeled using slopes and intercepts that were allowed to vary randomly between subjects, presented in Eq.

(2). The model included an overall intercept term (α_0) and listener-specific deflections from this intercept (L). The listener deflections from the intercept were constrained to have a mean of 0, and were modelled as coming from a normal distribution with a variance of σ^2_L. The VTL differences (ΔVTL), formant-differences (ΔF1, ΔF2, ΔF3), and interactions (ΔF1F2, ΔF2F3) were all included as continuous predictors. A binary variable 'Different' was also included in the model, which indicated whether the trial involved vowels of different categories (1) or not (0). This additional parameter allows for VTL differences (ΔVTL) to affect judgments in different ways in same- and different-phoneme trials; β_{VTLS} represents the estimated effect for ΔVTL in situations where the same vowel is presented, while $\beta_{VTLS} + \beta_{VTLD}$ is the estimated effect for ΔVTL in cases where different vowels are presented.

Each of the seven slope terms in Eq. (2) was broken down in an ANOVA-style decomposition as in Eq. (3). This decomposition models each of the beta terms in Eq. (2) as varying as a function of listener (β^L_k), training group (β^T_k), and a slope-specific mean term (β^0_k). The β^0_k terms for each predictor indicates the mean value of an effect across all groups and listeners, while the sum $\beta^0_k + \beta^T_k$ will indicate the strength of the effect for an individual group. The sum of the three coefficients in Eq. (3) will yield the estimated regression coefficient for a particular participant within a given training group. For example, β^0_{VTLS} gives the mean effect for ΔVTL (in same-phoneme trials) across all listeners, $\beta^0_{VTLS} + \beta^{T=1}_{VTLS}$ gives the mean value of this predictor for all listeners in the trained group, and $\beta^0_{VTLS} + \beta^{T=1}_{VTLS} + \beta^{L=4}_{VTLS}$ gives the value of this predictor for only one listener (Subject 4) in the trained group.

$$\beta_k = \beta^0_k + \beta^T_k + \beta^L_k \quad (3)$$

In order to make the parameter estimates in Eq. (3) identifiable, the deflections associated with each of the effects in Eq. (3) were constrained to sum to zero around the appropriate slope mean term (β^0_k), as described in Kruschke (2014, chap. 20). The listener-specific slope terms (β^L_k) were modelled as coming from normal distributions with a mean of zero and a parameter-specific variance term ($\sigma^2_{\beta^L_k}$). Because each β^T_k term only consisted of a single degree of freedom, the untrained group effect for each slope was fixed at 0, and the trained group effect ($\beta^{T=1}_k$) was estimated using a diffuse prior with a mean of 0 and a variance of 100. Estimated effects were then centered after estimation. The means for each bundle of regression coefficients (β^0_k for each slope term), and the overall intercept (α_0), were given a diffuse prior with a mean of 0 and a variance of 100. Each of the higher-population variance parameters ($\sigma^2_{\beta^L_k}$ for each slope term, and σ^2_L) were given a half-cauchy prior with a standard deviation of 3.

In all, 444 parameters were estimated from that data, on 15,288 total observations. Posterior samples for all parameters were generated using JAGS (Plummer et al., 2003; R Core Team, 2015). Four independent chains were run, with each chain being a total of 2,500 steps in length, for a total of 10,000 steps. A 10,000 step adaptation and burn-in was used, after which chains were thinned every 300th step to reduce autocorrelation in the chains and to maintain a reasonable file size. The chains mixed well, with the effective sample sizes of all parameters being nearly 10,000.

2.5.1.1. Predictions regarding directions of effects. Relative-size judgments were reported on a sliding scale where a positive value indicates that the first voice is larger and a negative value indicates that the second voice is larger. Since the FF and VTL difference predictors are calculated with respect to the first voice, a negative value indicates that the first voice has lower frequencies for the predictor of interest. As a result, an overall negative association is expected between all of the continuous predictors (ΔVTL, ΔF1, ΔF2, ΔF3) and perceived relative size differences. In other words, negative FF differences lead to positive size differences since larger speakers produce lower frequencies in general.

When the same category is being compared across the two vowels in the pair, all of the formant-difference predictors (ΔF1, ΔF2, ΔF3) will equal zero, and the indicator variable 'Different' will also equal 0. As a result, relative-size judgments will be explainable solely on the basis of ΔVTL as mediated by the VTL slope coefficient (β_{VTLS}). Since ΔVTL will be negative in cases where the first voice has a longer implied VTL, this coefficient is expected to be negative as long as listeners associate lower FFs (negative ΔVTL) with larger speakers in same-phoneme trials. In different-phoneme trials, the 'Different' coefficient will equal 1 and so β_{VTLD} will be estimated. This coefficient will reflect the difference in the effect for ΔVTL between same- and different-phoneme trials. Consequently, a non-zero value for β_{VTLD} would indicate that VTL information has different effects in different-phoneme trials relative to same-phoneme trials. It is worth noting that if relative speaker-size were solely determined by a reasonably-accurate VTL estimate in all situations, then the effect for ΔVTL on relative-size judgments should not be affected by whether the trial contains different phonemes or not, and so β_{VTLD} should equal zero.

In different-phoneme trials, the FF-difference predictors may have non-zero values since there will be at least some inherent (within-speaker) differences in formant patterns across the vowels. Consistent phoneme-biases in size judgments will be reflected as non-zero values for the formant-difference coefficients ($\beta_{\Delta F1}, \beta_{\Delta F2}, \beta_{\Delta F3}$). Just as with ΔVTL, the individual formant difference predictors are expected to be negative in the event that lower FFs for the first voice are associated with larger sizes for the first voice. Based on the results presented in Barreda (2016), it is expected that listeners will be biased towards identifying vowels with inherently-lower FFs as larger, independently of VTL differences between the voices. For example, /u/ has a much lower inherent F1 than /ɑ/, and this difference will be reflected by the value of ΔF1 between the vowels. As a result, a consistent bias in relative size-judgments between /u/ and /ɑ/ will be reflected by the value of $\beta_{\Delta F1}$. On the other hand, if relative-size judgments are solely based on the VTL differences between voices and there are no phoneme-biases, all formant difference predictors should equal 0, and should certainly not have large enough magnitudes such that they may overwhelm the effects for VTL differences between voices. The interaction terms will reflect the extent to which formant differences have non -additive effects on relative-size judgments. If combined formant-difference effects are smaller than expected (relative to these difference effects in

isolation), the interaction terms ($\beta_{\Delta F1}, \beta_{\Delta F2}$) will be positive, negating the negative relationship between formant differences and relative-size judgments. On the other hand, credibly negative interaction terms would indicate that combined formant-differences tend to have a larger than expected effect relative to independent differences in formants.

The group trained to report VTL will be examined to look for differences in comparison to the untrained group, which reported speaker height. Based on the results in Barreda (2016), it is expected that listeners in the untrained group will exhibit phoneme-biases, and that this will result in non-trivial values for $\beta_{\Delta F1}, \beta_{\Delta F2}$ and $\beta_{\Delta F3}$. In contrast, to the extent that listeners in the training group are reporting VTL, these formant-difference predictors should not be different from zero. In the event that these are non-zero, they may show diminished magnitudes relative to those of the untrained group. Relatedly, the trained group may exhibit increased magnitudes for the VTL predictors in the model ($\beta_{\Delta VTLS}, \beta_{\Delta VTLD}$) either because of an increased ability to perceive or report VTL, or because of a decreased reliance on phoneme-biases. On the other hand, if the pattern of effects is largely the same for both groups of listeners, this would be evidence of the fact that listeners either do not have ready access to VTL estimates that are independent of vowel quality, or that they are not able to report these estimates easily and with accuracy.

3. Results

A total of 15,288 responses were collected across the 46 subjects. Since the experiment was self-paced with a one-hour time limit, not all listeners completed all six blocks; an average of 332 responses per subject were collected. Listeners in the untrained group reported relative height differences, while listeners in the trained group reported 'voice size' (i.e., VTL) differences. Unless a distinction needs to be made between these judgments, for the presentation of the results and analysis these judgments will collectively be referred to as relative-size judgments. Training results are discussed in detail in the Appendix A. In general, listeners were quite accurate in identifying voice VTL, and demonstrated improvement throughout their training. However, there was also quite a bit of variability in performance.

In same-phoneme trials (Fig. 3a), perceived relative-size differences vary systematically as a linear function of the VTL difference between the voices. As expected, negative ΔVTL values were associated with larger apparent speakers and positive ΔVTL values were associated with smaller speakers. Importantly, the linear relationship between ΔVTL and perceived relative-size has a *y*-axis intercept near zero. This indicates that when the difference between the vocal tract lengths (ΔVTL) is near zero, the perceived relative-size difference is also near zero, meaning that speakers with roughly equal VTLs were judged to be of roughly equal size. When considered in isolation, these results support the hypothesis that listeners base relative size judgments on speaker VTL estimates. However, the results in different-phoneme trials (Fig. 3b–f) show quite a different pattern.

In the absence of phoneme-biases in size perception, all panels in Fig. 3 should look more or less like 3a. Instead, in addition to the linear relationship between ΔVTL and

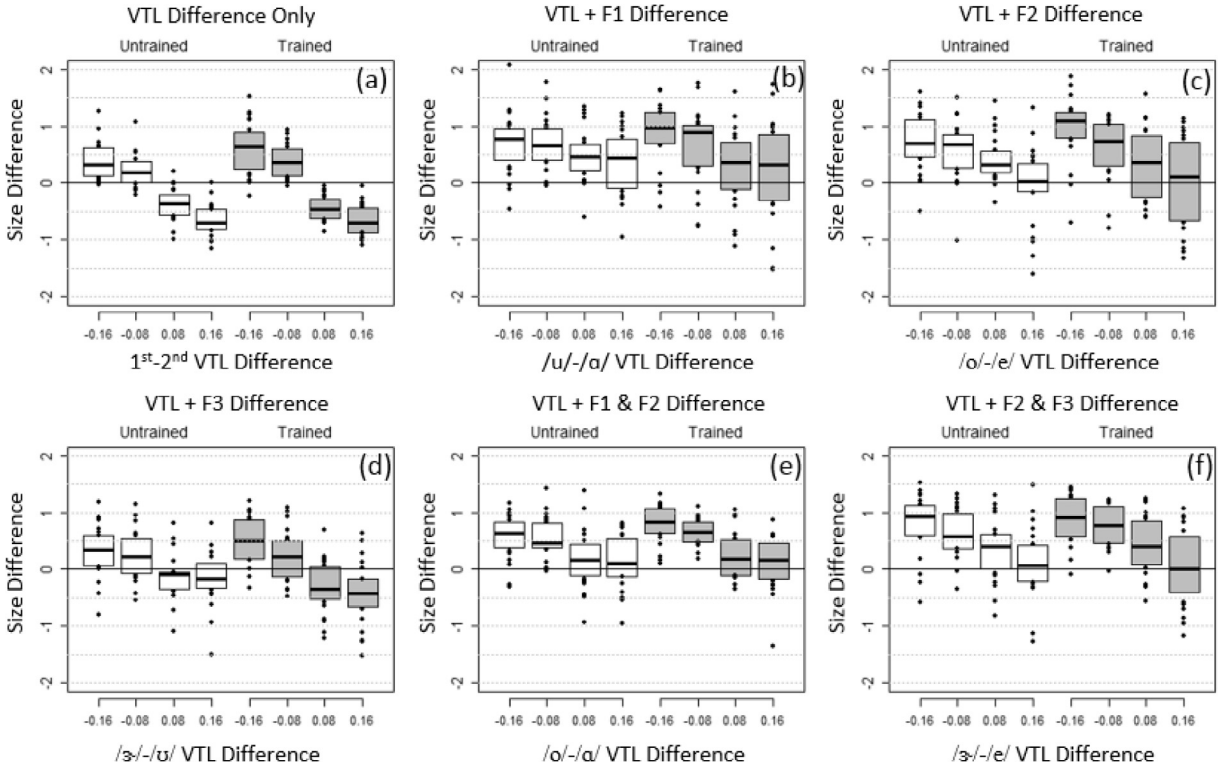

Fig. 3. Average relative-size judgments, within-listener, presented across different vowel pairs and VTL differences between the voices. VTL differences are quantified using the difference in scaling in log-Hz between the voices. A negative VTL difference indicates a longer VTL for the reference vowel. (a) Same-phoneme trials, where VTL and size differences are presented with respect to the first vowel in the pair. (b–f) Different-phoneme trials where VTL and size differences are presented with respect to a reference-vowel chosen for each pair. The reference vowel in each pair is the first vowel indicated below each panel.

relative-size judgments, most panels also feature an upward shift of this relationship that is consistent with phoneme-biases in size perception. Unlike in Fig. 3a, the relationship between ΔVTL and relative-size judgments in most panels in Fig. 3 do not have *y*-axis intercepts near zero. This means that for different-phoneme comparisons, when ΔVTL is near zero the perceived size difference between speakers will tend to have a positive value. In other words, listeners exhibited a tendency to identify the reference vowel chosen for each comparison in Fig. 3 as larger, independently of the ΔVTL difference between the voices. Since reference vowels were chosen on the basis of having inherently-lower FFs, the results in Fig. 3b-f demonstrate a bias towards identifying vowels with inherently-lower FFs as larger, independently of VTL information. For example, /o/ was chosen as the reference for the /o/-/e/ vowel pair (lower F2). In Fig. 3c we see that the perceived size difference between these two vowels varies as a function of the VTL difference between the speakers. However, /o/ tended to be identified as larger than /e/ even when presented with a substantially shorter VTL (positive ΔVTL values). This suggests that listeners may be associating the substantially lower F2 in /o/ (see Fig. 2c) with a larger speaker, even though the large difference in F2 is reflective of phonemic differences rather than differences in speaker size.

The results presented in Fig. 3 suggest that speaker-size judgments are affected by the particular spectral content of speech sounds, in addition to the apparent VTL difference between the speakers. The effects of VTL and formant differences on relative-size judgments will be investigated using the model outlined in Section 2.4. Table 2 and Fig. 4 present the means and 95% highest-density intervals (HDIs) of coefficient estimates for different effects, for trained and untrained listeners. The 95% highest-density interval indicates the range of 95% of the posterior distribution of a parameter (or a combination of parameters) such that every point inside the range is more probable than every point outside the range (Kruschke, 2010b). The 95% HDI can be used to establish a credible range of value for parameters, or linear combinations of parameters. If a value of interest (e.g., 0) is not within the 95% HDI of a parameter it is not a credible value for that parameter.

As seen in Fig. 4, individual differences across any of the first three formants have credible effects on perceived speaker-size, independently of the VTL difference between the vowels. The effects for the first two formants are roughly of the same magnitude, though the effect for differences in F3 is somewhat weaker. For untrained listeners both interaction terms are credibly positive, though the interval for the ΔF2F3 interaction contains values very close to zero. These results indicate that the effect of formant differences tend to be diminished when combined for untrained listeners. In contrast, for trained listeners the credible intervals of both interaction terms include zero, indicating that there is not good evidence for interactions between any of the formants.

VTL differences also had strong effects on perceived speaker-size, although effects were considerably weaker in different-phoneme trials. Given that the VTL differences implied by the stimulus vowels are equal in same- and different-phoneme trials, this indicates that listeners are better able to asses VTL differences between speakers when formant-pattern is held constant across the stimuli. The strength of effects for VTL differences might appear to be quite

Table 2
Mean values and lower and upper bounds of 95% highest-density intervals (HDI) for the regression coefficients for each predictor included in the model. For a given predictor *k*, trained indicates the mean value for trained listeners ($\beta_k^0 + \beta_k^{T=1}$), untrained indicates the mean value for untrained listeners ($\beta_k^0 + (\beta_k^{T=0}$), and difference indicates the difference in values between untrained and trained listeners ($\beta_k^{T=0} - \beta_k^{T=1}$).

		ΔF1	ΔF2	ΔF3	ΔF1F2	ΔF2F3	ΔVTLD	ΔVTLS
Untrained	Mean	−0.67	−0.57	−0.45	0.89	0.35	−1.68	−3.42
	HDI	−0.74, −0.61	−0.63, −0.51	−0.58, −0.33	0.24, 1.53	0.00, 0.68	−1.90, −1.47	−3.67, −3.18
Trained	Mean	−0.63	−0.71	−0.34	−0.26	−0.16	−2.52	−4.27
	HDI	−0.69, −0.58	−0.76, −0.65	−0.45, −0.23	−0.87, 0.35	−0.50, 0.17	−2.72, −2.37	−4.51, −4.06
Difference	Mean	−0.04	0.13	−0.11	1.15	0.51	0.84	0.85
	HDI	−0.12, 0.04	0.05, 0.21	−0.28, 0.05	0.24, 2.01	0.06, 0.99	0.60, 1.12	0.52, 1.18

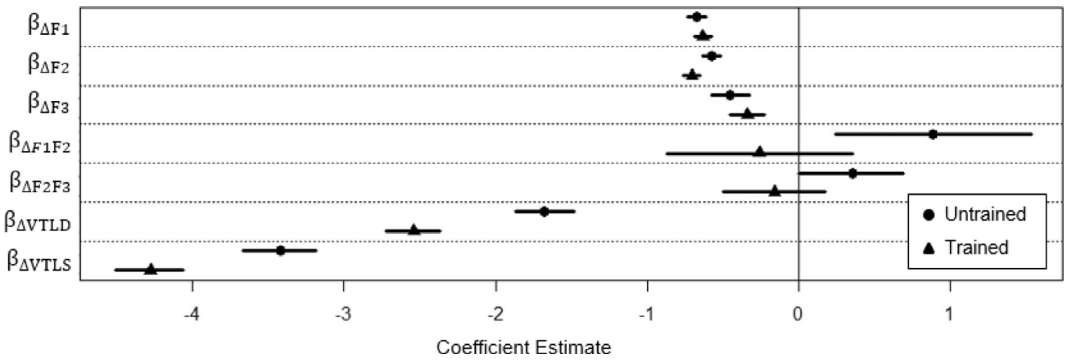

Fig. 4. Means and 95% highest-density intervals for slope coefficient estimates for different predictors, for trained and untrained listeners. Mean values for predictor k for trained listeners are equal to $\beta_k^0 + \beta_k^{T=1}$ and mean values for untrained listeners are equal to $\beta_k^0 + \beta_k^{T=0}$.

a bit larger than the effects for formant differences. However, this is tempered to some extent by the fact that inherent differences in the FFs across formants can be much larger than even the largest VTL differences between speakers. For example, the difference in F1 between /ɑ/ and /u/ is 0.85 log-Hz (log(646)–log(277)). Based on $\beta_{\Delta F1}$ for the untrained group presented in Table 2, the expected size-difference associated with this F1 difference is -0.57 (-0.67×0.85). In contrast, the expected size-difference associated with a 0.16 log-Hz VTL difference in different-phoneme trials for untrained listeners is only -0.27 (-1.68×0.16). In fact, we can see that most listeners identified /u/ as larger than /ɑ/ even when presented at a much shorter VTL (Fig. 3b). So, while the effect for VTL cues is much larger in magnitude than that of FF cues, because of the relatively small differences in VTL between speakers the effect of VTL information on size judgments can be smaller than that of phoneme-biases.

There are some differences in coefficient values between trained and untrained listeners in Fig. 4 (also presented in Table 2), particularly for the VTL effects and in the interaction terms. However, as might be expected given the similarity of the results presented in Fig. 3, the general pattern of coefficient values is broadly similar across the two groups. Both groups of listeners were strongly influenced by VTL differences in same- and different-phoneme trials. Although trained listeners had substantially larger effects for ΔVTL in both kinds of vowel pairs, they exhibited a large drop in the strength of the effect for ΔVTL in different-phoneme trials just as the untrained listeners did. Both sets of listeners also displayed credibly-negative effects for each formant-difference predictor ($\Delta F1, \Delta F2, \Delta F3$), indicating that both groups associate phonemically-determined (inherent) differences in FFs with larger speakers. Interestingly, the effects for inherent differences in FFs are not weaker for trained listeners, even though the effect for VTL is stronger. This indicates that although training listeners to report VTL may well increase their sensitivity to VTL cues, this information is used in addition to phoneme-biases instead of supplanting them.

3.1. Individual differences in the use of acoustic cues

As outlined in Section 2.5.1, the analysis included parameters that allowed for the effects of different predictors to vary randomly for each participant. This means that with the addition of the appropriate terms in Eq. (3), the beta term for any acoustic cue may be found for any particular listener. For example, the sum of $\beta_{F2}^{0} + \beta_{F2}^{T=0} + \beta_{F2}^{L=7}$ will yield the regression coefficient associated with ΔF2 for the seventh listener in the untrained group. Using this approach, the 46 listener-specific coefficient estimates for each of the main-effect predictors included in the model may be recovered.[2] We may use these listener-specific regression coefficients to examine individual differences in the use of acoustic cues (cue-weighting) in relative speaker-size judgments, as in Fig. 5.

The distribution of coefficients in the panels of Fig. 5 reveals that there is a large amount of between-listener differences in the use of acoustic cues. In fact, individual differences between-speakers within-groups are generally large in magnitude relative to the differences between trained and untrained listeners. The large amount of variation in cue-weighting between listeners suggests that there may be several different strategies for assessing speaker-size given the same spectral content in different-phoneme trials. For example, listeners toward the left of each column in Fig. 5 are those who were strongly influenced by VTL cues in different-phoneme trials. Many of these listeners show weak effects for differences in individual formants, indicating that these listeners may have actually been reporting VTL estimates that were relatively free of phoneme biases. On the other hand, listeners towards the right of each column were not strongly influenced by the VTL differences between the voices, and inherent differences between formants may have played a stronger role in their assessments of speaker size.

Fig. 6 presents all responses from three individual listeners in the untrained group, presented in a similar way as the overall data in Fig. 3. The individual model coefficients for these three listeners (U1, U3, U19) can be found on Fig. 5. These three listeners exhibit quite different approaches to the determination of speaker size, and reflect different levels of reliance on VTL cues and phoneme-specific spectral information. Listener U19 shows the least orderly responses in same-phoneme trials, however responses are still influenced to some degree by the VTL differences between the voices. In different-phoneme trials, there is not a strong linear relationship between VTL and relative-size judgments for this listener. Instead, this listener is strongly influenced by phoneme-biases and simply selects the vowel with the lower inherent FFs as larger in nearly all cases. This results in responses that are 'flat' across the VTL difference between the voices and are simply shifted upwards by the phoneme bias. In contrast, listener U1 appears to be responding primarily to the VTL-differences between vowels, even in different-phoneme trials. The relative lack of phoneme biases in this listener's judgments is reflected in the similarity of the responses in same- and different-phoneme trials. Finally, listener U3 exhibits a hybrid strategy whereby judgments are influenced both by the VTL differences between the voices and by phoneme biases. In different-phoneme trials this listener displays both the linear relationship between VTL and size differences, and the intercept shifts in these relationships resulting from phoneme-biases.

The different behaviors outlined above are reflected in the relative weights of each listener's coefficients in Fig. 5. For example, listener U19 has small VTL coefficients and large formant-difference coefficients, listener U1 has strong VTL coefficients and weak formant coefficients, and listener U3 has strong VTL and formant coefficients. By extension, the variety in absolute and relative weights afforded to different acoustic cues by each listener in Fig. 5 reflects the wide variety of strategies employed by different listeners. In general, listeners differ in the magnitude of the effects for different predictors, with almost all being negative. This indicates a general

[2] In fact, the coefficients resulting from this are nearly identical to those that are found by fitting the model outlined in Eq. (2) (minus the listener effect) to the data from each listener. The multilevel Bayesian model employed here was preferred to this 'no-pooling' approach for several reasons, including the fact this approach simultaneously estimates credible intervals for all parameters as seen in Fig. 5. For more information please see Section 2.4. The interaction terms are not included in this section because, since they were each estimated from a single vowel contrast, there is not much data to estimate these terms at the level of individual listeners. As a result, 46 of 46 listeners had estimate intervals that overlapped with zero for the F1 × F2 term and 39 of 46 listeners did so for the F2 × F3 term.

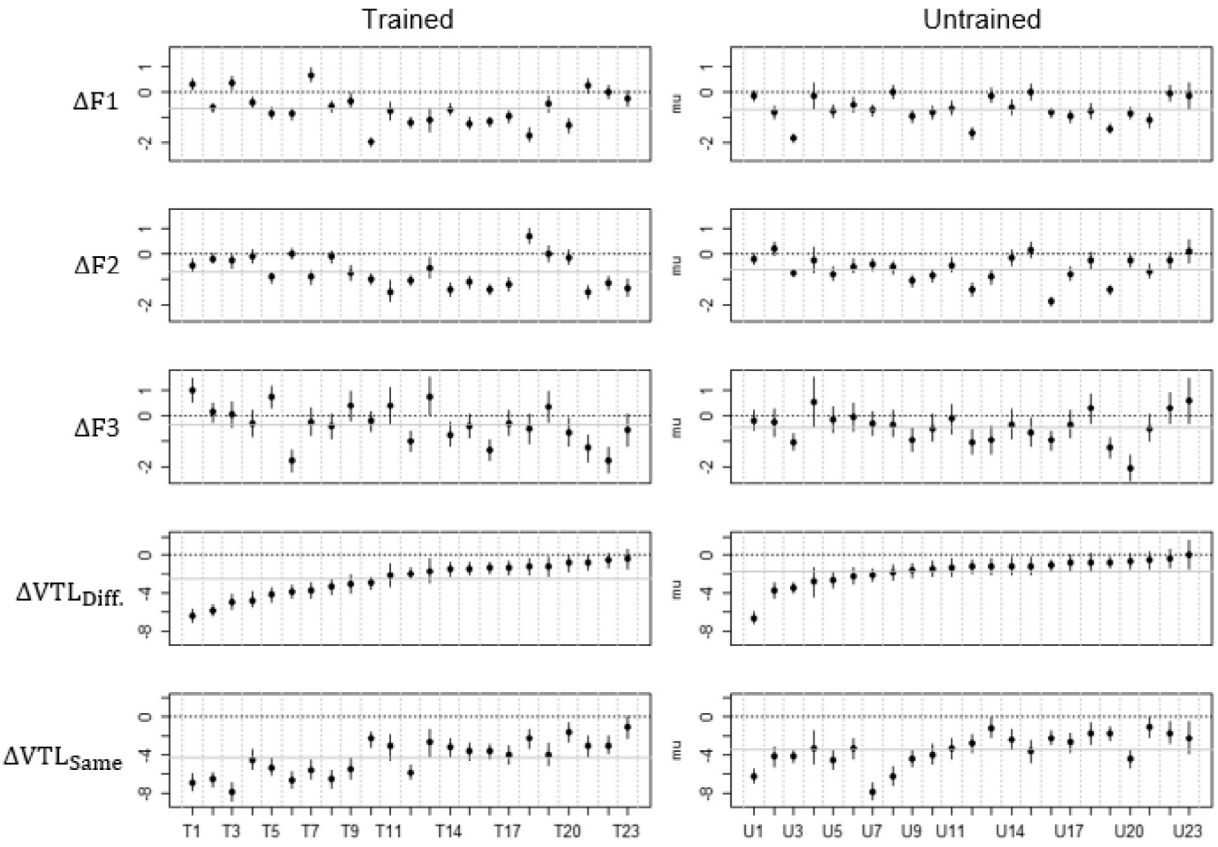

Fig. 5. Each panel shows listener-specific coefficient estimates for different effects (i.e., the sum of the appropriate $\beta_k^0 + \beta_k^T + \beta_k^L$ terms from Eq. (3)), and their 95% HDI. The left column features trained listeners while the right column features untrained listeners. Solid horizontal lines indicate group means. Within each column, listeners are sorted based on their mean effect magnitude for VTL in different-phoneme trials so that each column indicates coefficient values for a single listener. For example, the leftmost set of values inside each column correspond to the coefficient estimates for the listener with the largest VTL effect in different-phoneme trials, within each training group.

tendency to associate lower frequencies with larger speakers, with variation existing primarily in the extent to which a given cue is used, and if it is used at all.

4. Discussion

The experiment outlined above had two primary purposes. First, the experiment was meant to investigate phoneme-biases in size perception and the sensitivity of these biases to inherent differences in FFs across vowel phonemes. Second, the experiment was meant to investigate the availability of VTL information in size judgments by comparing results from an untrained group of listeners reporting speaker height to, VTL judgments from a group of listeners trained to report VTL directly.

4.1. Phoneme biases

Barreda (2016) reported varying biases towards identifying some vowels as larger than others, independently of VTL information, and suggested that this may be a result of differences in the inherent spectral characteristics of the vowels being compared. Based on the overall marginal association between larger speakers and lower FFs, vowels with inherently-lower FFs were expected to be associated with relatively-larger speakers, above and beyond any given VTL cues. The experiment outlined above confirms this expectation, and finds a bias towards identifying vowels with inherently-lower formants as larger in all of the different-phoneme vowel pairs considered. In addition, in most cases these phoneme biases were large enough to overwhelm even relatively large simulated VTL-differences between voices. Phoneme-biases were present in the height judgments made by nearly all listeners, and these were not reduced in the judgments of listeners trained to report VTL differences directly. Taken together with the results presented in Barreda (2016), these findings indicate that rather than only considering apparent VTL differences between voices, listeners also consider the specific spectral characteristics of the vowel sounds being compared when making relative-size judgments. This leads to a tendency to overestimate the size of speakers when they produce vowels with inherently-low FFs, independently of any VTL differences that may also exist between the voices.

These results have implications for the accurate modelling of judgments of speaker size, and for future investigations of size perception generally. Much investigation into the perception of speaker size has hinged on the idea that listeners extract VTL estimates in a relatively accurate and phoneme-independent manner, even from very short stretches of speech. For example, this assumption is made at least implicitly any time a researcher investigates size perception on the basis of monosyllabic words or isolated vowels, and considers that these limited stimuli reveal VTL information to the listener (Collins, 2000; Rendall et al., 2007; Smith & Patterson, 2005;

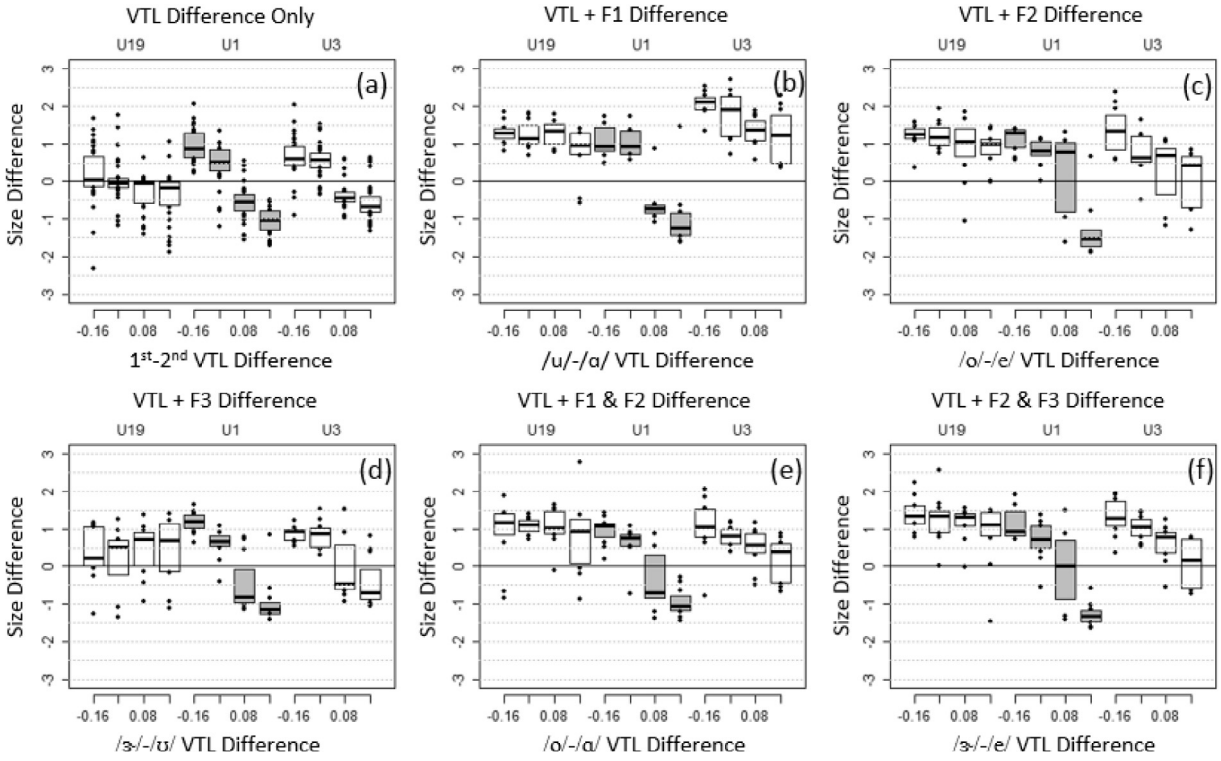

Fig. 6. Relative-size judgments, presented across different vowel pairs and VTL differences for selected untrained-listeners. A negative VTL difference is associated with a longer simulated VTL for the reference vowel in the pair. (a) Average relative-size judgments across all same-phoneme trials, where the first vowel in the pair acts as the reference. (b–f) Average relative-scale judgments for different-phoneme trials, where the reference vowel is the first vowel in the pair as indicated below each panel.

Smith et al., 2005). In this view, VTL information is immediately available for a speaker, and all speech produced by that speaker contains 'the same' spectral information (i.e., VTL). However, if listeners also consider the spectral content of sounds directly, then size information in the speech produced by a speaker may vary as a function of phonemic content. As a result, the perception of speaker-size should be modelled on the basis of specific spectral characteristics (i.e., the actual spectral content of the sound) and not simply on the basis of speaker VTL information, and investigated on a token-by-token basis rather than aggregated across all tokens produced for a speaker/voice. In general, considering only aggregate or VTL characteristics may obscure meaningful, systematic variability in size judgments as a result of the actual spectral content of a sound.

4.2. Availability of vocal-tract length estimates for size judgments

The availability of VTL information is assumed in much research on size perception, and some researchers have suggested that VTL estimates are automatically provided to listeners by the peripheral auditory system (Irino & Patterson, 2002; Ives et al., 2005; Patterson & Irino, 2014; Smith & Patterson, 2005; Smith et al., 2005; Turner et al., 2006). If speaker-size judgments were driven by VTL and listeners had easy access to this information, listeners should not require very much training in order to report it. However, very few listeners in either the trained or the untrained group were able to report VTL independently of phoneme-biases. Furthermore, the pattern of phoneme-biases is quite similar across both groups of listeners (Fig. 3), suggesting that even when explicitly familiarized with the acoustic characteristic associated with VTL differences between speakers (i.e., uniform scaling of formant patterns), listeners have difficulty reporting VTL independently of phoneme-specific formant-pattern information. This difficulty in reporting relative speaker-size independently of phoneme-specific spectral information suggests that either listeners do not have easy access to a speaker-VTL estimate, or that relative-size judgments are not only based on a comparison of these VTL estimates.

The large amount of individual differences in listening strategies is also problematic for theories that suggest that listeners have access to reliable VTL estimates. Every substantially-different configuration of the coefficients in Fig. 5 can be thought of as a different listening-strategy for estimating relative size from acoustics. For example, of the three listeners in Fig. 6, one relies mostly on phoneme-specific information, another mostly on VTL and another on both phoneme-information and VTL. If VTL estimates were the primary determinants of perceived relative-size in human listeners, it is not clear why some listeners would choose to ignore VTL information and be so influenced by phoneme-specific information. Further, if VTL estimates were available to all listeners due to automatic processing carried out by the peripheral auditory system, it is not really clear why such diverse (and in many cases suboptimal) listening strategies would arise.

The difference in the effects of VTL in same- and different phoneme trials also suggests that VTL estimates are not automatically and easily available to all listeners. For example, consider any two vowels presented in a pair, at a given VTL difference. The results presented in Fig. 4 indicate that the effect of the apparent VTL-differences between these vowels will

depend on whether the same vowel is being presented or not, with VTL having a substantially weaker effect in different-phoneme trials. If listeners could easily recover a phoneme-independent VTL estimate, VTL differences should not have a weaker effect on perceived size simply because two different vowels are being presented. However, as outlined in Section 1.3, different-phoneme trials actually require that listeners isolate VTL information from phoneme information in formant patterns when making size judgments. As a result, the weaker effect for VTL in different-phoneme trials suggests that VTL information may simply be more difficult to identify in different-phoneme trials, indicating that segregating size/VTL information from phonemically-dependent spectral information in formant patterns is not easy, automatic, or to be taken for granted.

The overall pattern of results presented here is problematic for some of the stronger claims made regarding the availability of a speaker-dependent VTL estimate that is independent of linguistically-determined spectral information. Although it may be the case that listeners estimate VTL during speech perception, their internal representations of these estimates may not be available to them in a manner that is directly comparable across different vowel qualities (i.e., different underlying spectral shapes). Rather than having easy access to VTL-estimates that are free of phoneme-biases, reporting VTL information in speech sounds appears to be a skill that can vary dramatically between listeners, and which can be improved with training. Although trained listeners did not have diminished phoneme-biases, they were more responsive to VTL differences in both same- and different-phoneme trials. As reported in the Appendix A, listeners showed a modest but significant improvement in their judgments even during the relatively brief training given to them. Using a different training method to teach listeners to report speaker VTL, Barreda and Nearey (2013b) similarly reported improved accuracy in VTL estimates after training, in addition to a significant advantage in accuracy for listeners that had received formal musical training at some point in their lives.

4.3. Suboptimal use of acoustic information in size estimation

Undoubtedly, a reliance on the absolute spectral characteristics of vowel sounds rather than solely on VTL estimates represents a suboptimal use of acoustic cues in size estimation. In general, responding to phoneme-specific spectral characteristics instead of a VTL estimate that is stable for a given speaker will result in more variability in the size judged for that speaker, and instability across utterances with different linguistic content. It bears noting however, that this poor use of acoustic information is very much in line with previous findings regarding suboptimal use of cues in how listeners estimate speaker size from speech. As noted in the introduction, human listeners are not very accurate at identifying the veridical size of adult speakers from acoustics. This may occur simply because random variability in speaker size overwhelms the systematic variability between VTL and height when restricted to adult VTL ranges (Barreda, 2016). However, above and beyond any inherent difficulties in accurately identifying speaker size from speech, it appears that human listeners do not behave in a manner that would lead to accurate estimation of speaker size even if reliable size cues were available. For example, speaking f0 is highly variable within speakers so that small differences in f0 offer absolutely no evidence regarding speaker size. Despite this, differences in f0 as small as 20 Hz can overwhelm relatively large VTL differences and people will identify the speaker with lower f0 as larger despite having a shorter apparent VTL (Rendall et al., 2007). This behavior has been noted by many researchers, for example Pisanski et al. (2014) state that "[i]n the absence of a strong physical relationship, the strong perceptual association between F0 and size poses a paradox" (95).

The strong influence of f0 on size judgments may appear paradoxical given the traditional perspective of size perception where spectral information is expected to affect perceived size by informing VTL judgments. In this view, VTL judgments represent 'corrected' spectral information that is expected to be true and stable for a speaker. If listeners were doing this for spectral information, we might also expect that some correction may be employed when using f0 in order to account for its large amount of variation within speakers. In general, we may consider that this would represent a better approach to the use of acoustic information in size judgments. Instead, the results presented here suggest that listeners do not base relative-size judgments solely on 'corrected' VTL cues, but also respond to phoneme-specific spectral content. Although this decidedly suboptimal strategy may seem puzzling in isolation, it is very much in line with the strong influence of f0 on size judgments. In general, it seems that the perception of relative speaker-size may operate at a relatively low level of processing, where many listeners appear to simply make a global association between low-frequency spectral energy, or a low f0, and larger speakers.

4.4. Future investigation of the perception of speaker size

Responding to implied VTL-differences independently of phoneme-specific information appears to be somewhat difficult for listeners. Listeners also differ substantially in their ability to respond to VTL independently of phoneme information, and there are many different strategies for estimating relative size from speech in different-phoneme trials. These results suggest that phoneme-independent VTL estimates may not be easily available to listeners, and that reporting VTL may be an ability that can differ substantially between listeners, and which can be improved with training. This suggests that researchers may need to consider the methods used to investigate the perception of speaker size, in order to get an accurate impression of how listeners arrive at speaker size judgments. For example, relative and absolute size-judgments might be carried out in different ways such that the direct comparison of spectral characteristics might have more influence in relative size judgments compared to absolute size judgments. When all size judgments, be they relative or absolute, are thought to arise solely from the consideration of VTL estimates, such task-specific differences in responses are not expected. Furthermore, speaker-size judgments should be investigated given the actual spectral content of the sound rather than using some speaker-dependent measure of VTL. Modelling size judgments on the basis of speaker VTL information while ignoring inherent FF differences across utterances with differing linguistic content may obscure true listener behavior.

In addition, further investigation into the specific use and integration of spectral information in speech sounds is warranted. In particular, the weak effects for F3 differences or perceived size differences, and the potential interaction effects between formant differences on size judgments suggest that the use of spectral information in size perception may be relatively complex. As seen in Figs. 3 and 4, the effect for F3 differences on size judgments were not as large as those of F1 and F2. It may be the case that rhotic vowels are not treated as fully vowel-like by listeners. For example, voiced stops such as /b/ contain mostly very low-frequency energy. However, it is not clear if listeners would conclude that speakers are extremely large when judging size solely on a produced /b/, or if some other mechanism is used for non-vocalic sounds. Another possibility is that the very low F3 of /ɝ/ is perceptually merged with F2, resulting in a single medium-frequency formant peak rather than a mid F2 and a very low F3 (Chistovich & Lublinskaya, 1979).

Similarly, the interactions between formant-differences for the untrained group seem to suggest that the use of spectral information is more complicated than what can be described by only considering formant differences. Although the interaction terms were not credibly different from zero for the trained group, this should in no way be interpreted as suggesting that there is a lack of interactions between formant differences in general. For example, it seems reasonable to expect that if all three formants were allowed to vary independently between two vowels in a pair, at least some important interactions between formant differences would be present. In other words, if listeners were asked to compare pairs of vowels that were randomly selected from a vowel space, it seems unlikely that these judgments could be fully explained on the basis of formant-difference terms with no role for interactions between the formants. In general, these results suggest that a direct comparison of formant differences, while a useful paradigm for the analysis of experimental results, should not be thought of as fully capturing the mapping between the spectral information in speech sounds and apparent speaker-size.

Finally, future investigation into the perception of speaker size might consider the time course of size estimation, and how initial estimates might be refined as more information becomes available. If one thinks that accurate VTL information is available to listeners from even very short stretches of speech, then one is essentially suggesting that size estimates should be relatively accurate from the start and so they should be relatively stable across time. However, if listeners also consider spectral content directly, this suggests that although phoneme-biases may result in large effects in size judgments initially, these may get smoothed-out as more information becomes available. For example, if listeners were comparing pairs of vowel pairs (e.g., / ɑ u / vs. / e o /) from voices with different VTLs, we might expect that the biases associated with each vowel would tend to cancel out, and listeners would be more likely to respond to the underlying VTL differences. In essence, if listeners are simply responding to direct spectral evidence they may simply be estimating something like a long-term average spectrum for speakers, which would effectively become a reasonably-accurate speaker-VTL estimate given enough exposure to a speaker. The possibility that speaker-size perception from spectral information involves aggregation across multiple speech sounds raises several interesting questions regarding the manner in which this information is aggregated, and the time course of this aggregation, that do not arise when VTL estimation is easy and accurate even from very short stretches of speech. However, it bears noting that even this aggregation across stimuli may not lead to accurate size judgments because of the noisy relationship between VTL cues and speaker size in adults (Collins, 2000; Rendall et al., 2007; Pisanski et al., 2014; Van Dommelen & Moxness, 1995).

5. Conclusion

The experiment outlined above confirms the presence of phoneme-biases in size perception, and indicates that these may have a large influence on judgments of relative speaker-size. Listeners are quite sensitive to inherent formant-frequency differences between vowel phonemes when making relative-size judgments: Phonemically-determined spectral differences triggered phoneme-biases in all of the different-phoneme pairs considered. Furthermore, results suggest that listeners cannot easily make relative-VTL judgments that are independent of phoneme-specific FF information. Relative-VTL judgments made by listeners trained to report VTL looked broadly similar to relative-height judgments made by untrained listeners. Both groups of listeners showed a strong influence for phoneme biases, and both had a weaker effect for VTL in different-phoneme trials.

Overall, results suggest that listeners do not appear to have easy access to phoneme-independent speaker-VTL estimates, or that they do not base relative-height judgments solely on such estimates. Instead, in addition to responding to VTL differences between voices, listeners respond to the phoneme-specific spectral content in speech sounds, associating inherently lower FFs with larger speakers independently of VTL information. Although this behavior certainly represents a suboptimal use of the spectral information in speech sounds, it is very much in line with the general suboptimal use of speech cues in size judgments. For example, listeners also rely heavily on f0, which is a very poor cue to speaker size in adults (within sex). Once a rigid attachment to VTL-based size perception is abandoned, many interesting questions arise regarding the manner in which spectral information is considered, both within-phoneme and across time.

Appendix A. Training results

The purpose of analyzing the training data is simply to give an idea of the accuracy and degree of improvement for the trained group of listeners. As a result, analysis of that data will be carried out using more traditional frequentist model-fitting and inference, including the ‘no-pooling’ approach of fitting models to the data obtained from each listener, and carrying out analyses on the distribution of estimates coefficients across all listeners (Gumpertz & Pantula, 1989; Lorch & Myers, 1990). Performance will be quantified based on absolute error in VTL identification. Voices were spaced on the board such that a voice 20 pixels to the right of another had FFs that were greater by a multiplicative factor of 1.012 (1.2%). Because of this relationship, we may associate 20 pix-

els with an increase in the FFs of 0.01193 log-Hz (log(1.012)), and a difference of a single pixel with an increase of 0.00059643 log-Hz. Based on this relationship, errors will be quantified by finding the error (or average error) in pixels, converting this value into log-Hz and then exponentiating this value to yield a measure of the proportional error. For example, clicking on a location 40 pixels away from the true location of a voice would indicate that a listener has overestimated voice VTL by 2.4% ($\exp(0.00059643 \times 40)$).

Although all listeners in the training group performed the same length training, since the training was self-paced subjects carried out between 50 and 147 trials (mean 89). The mean absolute error for first guesses across all listeners was 12.0% (min = 7.1%, max = 17.3%), and 10.0% (min = 6.6%, max = 14.0%) for second guesses. However, these values may underestimate accuracy because many listeners had small numbers of large errors. The median absolute error, within-listener, was 9.5% (min = 4.9%, max = 14.9%) for first guesses and 7.4% (min = 4.7%, max = 10.3%) for second guesses. Listeners showed a tendency to improve from their first guess to their second guess, with an average decrease in error of 1.8% (min = 0.24%, max = 5.67%), with every single listener showing a reduction in the VTL reporting error. Within-listener average VTL errors across first and second guesses are presented in Fig. A1.

These results indicate that listeners are reasonably accurate when reporting VTL in conditions where phonetic content is fixed (as in same-phoneme trials in the main experiment). Although an average difference of 1.8% across first and second guesses may not appear large, it represents a reduction of 15% in VTL estimation error relative to the average first guess (12% error). Furthermore, this reduction brings the average error down to 10% which is not far from the estimated JNDs for VTL in isolated vowels. Keeping in mind that listeners had to report VTL absolutely and independently of f0 information, making the task substantially more difficult than a two-alternative forced-choice task, this magnitude of VTL reporting error is quite good. In addition, the fact that listeners improved from their first to their second guess indicates that the auditory feedback provided after their initial guess allowed them to refine and improve their VTL estimates further.

In order to look for improvement during the training, a regression model was fit to absolute VTL errors individually for each listener, with the single predictor being normalized trial number. Trial number was normalized by subtracting one from the trial number and dividing this by the number of trials minus one. The result of this is that the first trial will be equal to zero and the final trial will be equal to 1. As a result of this coding, when regressing absolute VTL error on normalized trial number the intercept provides an estimate of initial accuracy, the slope coefficient denotes improvement during the course of the training, and the sum of the slope and intercept coefficient can be used to estimate final accuracy. This analysis was carried out for each listener individually and independently for first (Fig. A1b) and second guesses (Fig. A1c).

Average estimated initial accuracy for first guesses was 13.2% across all listeners (min 7.6%, max = 22.6%) and listeners demonstrated an average improvement of 2.1% (min = −6.0%, max 8.4%) which was significantly different from zero ($t(22) = 2.3$, $p = 0.029$). The same general pattern was evident for second guesses: average estimated initial accuracy for second guesses was 11.1% across all listeners (min 6.6%, max = 17.4%) and listeners demonstrated an average improvement of 1.74% (min = −8.0%, max = 6.8%) which was significantly different from zero ($t(22) = 2.4$, $p = 0.027$). The ranges for improvement include negative values, indicating that some listeners actually got worse as the training went on. This result is clearly evident in Fig. A1b and c where some of the arrows point upwards. However, we may recall no listeners did worse (on average) for second guesses relative to first guesses. This suggests that some listeners may have performed more poorly as the training went on due to fatigue or lack of concentration, rather than because they did not understand the task. This notion is reinforced by the strong negative correlations between initial accuracy (intercepts) and improvement (slopes) within listeners for the models examining improvement within first ($r = -0.78$) and second guesses ($r = -0.75$) across the training. Basically, listeners who were least accurate initially showed the most improvement during training and those who were most accurate had the least improvement (or got worse). We may also note that the listeners who showed decreased performance as the training went on were among the most accurate initially. This suggests that, for these listeners at least, accurate VTL estimation may have been costly in terms of attention or cognitive processing such that a continued effort of the kind required to perform as accurately as they did initially is quite difficult to maintain for an extended period of time.

An exploratory analysis revealed no relationship between sensitivity to VTL cues (based on magnitudes of ΔVTL effects

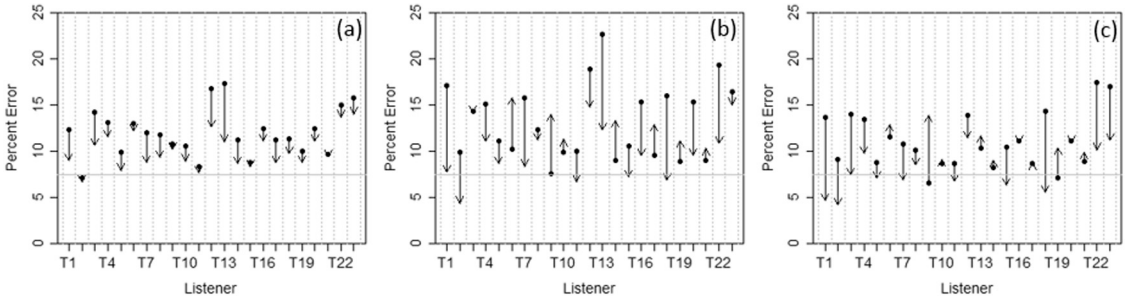

Fig. A1. (a) Points indicate average within-listener VTL error for first guesses, arrows indicate the magnitude of this error for second guesses. (b) Points indicate estimated initial VTL error for first guesses, arrows indicate estimated final error for first guesses. (c) Points indicate estimated initial VTL error for second guesses, arrows indicate estimated final error for second guesses. Values are presented using the same ordering of listeners as in Fig. 5, based on sensitivity to VTL cues in different-phoneme trials. Horizontal lines indicated just-noticeable differences for VTL in isolated vowels as reported by Smith et al. (2005).

for each listener, presented in Fig. 5) and training performance as measured by the statistics presented above. For example, within-speaker average errors presented in Fig. A1 are ordered on the basis of listener sensitivity to VTL in different-phoneme trials, just as in Fig. 5, and no clear pattern is evident. Given the amount of variability in performance, both between listeners and within-listeners across time, it may be the case that the relationship between performances on these two relatively complex tasks is too weak or noisy to observe given the limited number of subjects in the training group.

A.1. Summary

Overall listener performance was quite good, and many listeners performed at a level very near to estimated JNDs for VTL differences (4–8%) on what is a relatively difficult task (absolute reporting of f0 and VTL independently). Listeners show both local and global improvement in their guesses: their second guess tended to be more accurate than their first guess, and both first and second guesses improved as the training went on. In addition, although some listeners were quite accurate, there was a large amount of variability between listeners, and some initially accurate listeners exhibited decreased performance as the task went on.

References

Barreda, S. (2013). *Cognitively-active speaker normalization based on formant-frequency scaling estimation* PhD thesis. University of Alberta.

Barreda, S. (2016). Investigating the use of formant frequencies in listener judgments of speaker size. *Journal of Phonetics, 55*, 1–18.

Barreda, S., & Nearey, T. M. (2012). The direct and indirect roles of fundamental frequency in vowel perception. *Journal of the Acoustical Society of America, 131*, 466–477.

Barreda, S., & Nearey, T. M. (2013a). The perception of formant-frequency range is affected by veridical and judged fundamental frequency. *Journal of the Acoustical Society of America, 133*, 3567. http://dx.doi.org/10.1121/1.4806520. 3567.

Barreda, S., & Nearey, T. (2013b). Training listeners to report the acoustic correlate of formant-frequency scaling using synthetic voices. *Journal of the Acoustical Society of America, 133*(2), 1065–1077.

Bruckert, L., Liénard, J.-S., Lacroix, A., Kreutzer, M., & Leboucher, G. (2006). Women use voice parameters to assess men's characteristics. *Proceedings of the Royal Society B: Biological Sciences, 273*, 83–89. http://dx.doi.org/10.1098/rspb.2005.3265.

Chistovich, L. A., & Lublinskaya, V. V. (1979). The 'center of gravity'effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing Research, 1*, 185–195.

Collins, S. A. (2000). Men's voices and women's choices. *Animal Behaviour, 60*, 773–780. http://dx.doi.org/10.1006/anbe.2000.1523.

Core Team, R. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

Fant, G. (1970). *Acoustic theory of speech production*. Walter de Gruyter. 344 pages.

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

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

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*, 189–211.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251.

González, J. (2004). Formant frequencies and body size of speaker: A weak relationship in adult humans. *Journal of Phonetics, 32*, 277–287. http://dx.doi.org/10.1016/S0095-4470(03)00049-4.

Gumpertz, M., & Pantula, S. G. (1989). A simple approach to inference in random coefficient models. *American Statistician, 43*, 203–210. http://dx.doi.org/10.2307/2685362.

Hillenbrand, J. M., & Clark, M. J. (2009). The role of f0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics, 71*, 1150–1166. http://dx.doi.org/10.3758/APP.71.5.1150.

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

Hollien, H., Green, R., & Massey, K. (1994). Longitudinal research on adolescent voice change in males. *Journal of the Acoustical Society of America, 96*, 2646–2654. http://dx.doi.org/10.1121/1.411275.

Holmes, J. N. (1983). Formant synthesizers: Cascade or parallel? *Speech Communication, 2*, 251–273. http://dx.doi.org/10.1016/0167-6393(83)90044-4.

Irino, T., & Patterson, R. D. (2002). Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform. *Speech Communication, 36*, 181–203. http://dx.doi.org/10.1016/S0167-6393(00)00085-6.

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

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

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

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America, 67*, 971–995.

Kruschke, J. (2010). *Doing Bayesian data analysis: A tutorial introduction with R*. Academic Press.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences, 14*, 293–300. http://dx.doi.org/10.1016/j.tics.2010.05.001.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods, 15*, 722–752.

Kruschke, J. K., & Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. *The Oxford Handbook of Computational and Mathematical Psychology*.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America, 29*, 98–104. http://dx.doi.org/10.1121/1.1908694.

Lass, N. J., & Brown, W. S. (1978). Correlational study of speakers' heights, weights, body surface areas, and speaking fundamental frequencies. *Journal of the Acoustical Society of America, 63*, 1218–1220. http://dx.doi.org/10.1121/1.381808.

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 16*, 149–157. http://dx.doi.org/10.1037/0278-7393.16.1.149.

Nearey, T. M. (1978). *Phonetic feature systems for vowels*. Indiana University Linguistics Club.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America, 85*, 2088–2113.

Nearey, T. M., & Assmann, P. F. (2007). Probabilistic 'sliding-template'models for indirect vowel normalization. In *Experimental Approches to Phonology*. .

Patterson, R. D., & Irino, T. (2014). Size matters in hearing: How the auditory system normalizes the sounds of speech and music for source size. In *Perspectives on auditory research* (pp. 417–440). Springer.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*, 175–184. http://dx.doi.org/10.1121/1.1906875.

Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J. M., Röder, S., Andrews, P. W., et al. (2014). Vocal indicators of body size in men and women: A meta-analysis. *Animal Behaviour, 95*, 89–99. http://dx.doi.org/10.1016/j.anbehav.2014.06.011.

Plummer, M. et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proc. 3rd Int. Workshop Distrib. Stat. Comput.* (pp. 125). Technische Universit at Wien.

Podhorski, A. (1998). *Helium speech normalisation using analysis-synthesis method with separate processing of spectral envelope and fundamental frequency* PhD thesis. Faculty of Electrical Engineering, Technical University of Szczecin.

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

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

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

Smith, D. R. R., Walters, T. C., & Patterson, R. D. (2007). Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled. *Journal of the Acoustical Society of America, 122*, 3628. http://dx.doi.org/10.1121/1.2799507.

Turner, R. E., Al-Hames, M. A., Smith, D. R., Kawahara, H., Irino, T., & Patterson, R. D. (2006). Vowel normalisation: Time-domain processing of the internal dynamics of speech. In P. Divenyi (Ed.), *Dynamics of speech production and perception*. Amsterdam: IOS Press.

Turner, R. E., Walters, T. C., Monaghan, J. J. M., & Patterson, R. D. (2009). A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data. *Journal of the Acoustical Society of America, 125*, 2374. http://dx.doi.org/10.1121/1.3079772.

Van Dommelen, W. A., & Moxness, B. H. (1995). Acoustic parameters in speaker height and weight identification: Sex-specific behaviour. *Language and Speech, 38*, 267–287. http://dx.doi.org/10.1177/002383099503800304.