Journal of Phonetics 55 (2016) 1–18

Contents lists available at ScienceDirect

Journal of Phonetics

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

Research Article

Investigating the use of formant frequencies in listener judgments of speaker size

Santiago Barreda *

University of California, Davis, Department of Linguistics, 469 Kerr Hall, One Shields Avenue, Davis, CA 95616, USA

ARTICLE INFO

Article history:
Received 6 December 2014
Received in revised form
20 November 2015
Accepted 24 November 2015

Keywords:
Vowel perception
Speaker size perception
Speaker characteristics
Speaker normalization
Higher formants

ABSTRACT

The formant-pattern present in a given vowel sound will be determined by the vocal-tract length (VTL) of the speaker as well as by phoneme-specific information. Although human listeners tend to associate lower formant-frequencies with larger speakers, it is unclear whether they are responding to VTL information in speech sounds, or simply responding to the formant-pattern present in the sound. In this experiment listeners were presented with pairs of synthetic vowels from the set of (/i æ ʊ/), which could differ on the basis of simulated VTL and vowel category, within-pair. Listeners were divided into groups based on the number of formants contained by stimulus vowels (2, 3, 4, and 5-formant vowel groups). For each trial, listeners were asked to indicate which vowel sounded like it had been produced by a taller speaker. Results indicate that listeners do not rely solely on VTL cues when making speaker-size judgments, and that they exhibit biases towards selecting given phonemes as taller, even when contrary to the VTL differences between the voices. Furthermore, the higher formants (up to *F*5) are used by listeners when making speaker-size judgments, though not in a manner consistent with VTL-based speaker-size judgments.

1. Introduction

It has long been noted that the human voice carries indexical information about the physical and social characteristics of the speaker, in addition to conveying linguistic information (Labov, 1972; Ladefoged & Broadbent, 1957). The two most important cues to speaker size and gender are speaking fundamental frequency (*f*0) and the spectral characteristics of the speaker's voice, typically discussed using the formant frequencies (see González, 2006, for a review). Although *f*0 is largely under the control of the speaker, a speaker's mean *f*0 will largely be determined by the length and mass of their vocal folds (Titze, 1989). The range of formant frequencies (FFs) produced by a speaker, and the FFs typical for a given phoneme of the speaker's language, will be most strongly determined by the speaker's vocal-tract length (VTL). In general, speakers with longer vocal-tracts produce lower formants overall than speakers with shorter vocal tracts (Fant, 1970). As a result of this, when the entire human population is considered, larger speakers tend to produce speech with lower *f*0 and FFs overall than smaller speakers (Hillenbrand, Getty, Clark, & Wheeler, 1995; Peterson & Barney, 1952).

Although average *f*0 and FFs vary systematically between age and sex categories, the degree of systematicity between body size and speech acoustics appears to be much less consistent when one controls for sex and age (González, 2004; Hollien, Green, & Massey, 1994; Lass & Brown, 1978; Rendall, Vokey, & Nemeth, 2007; Van Dommelen & Moxness, 1995). A metastudy by Pisanski et al. (2014) compared the results of 39 independent datasets reporting correlations between acoustic voice parameters and body size measurements, comprising observations from over 1000 adult speakers. The authors find that while there is a statistically significant correlation between adult speaker-size and the acoustic parameters of their voices such as *f*0 or the average FFs, a large number of observations (618 men and 2140 women for *f*0, 99 men and 164 women for FFs) may be necessary in a given dataset in order to have the statistical power to identify the relationship.

The weakness of the relationship between speaker size and voice parameters in adults is caused, at least in part, by the restriction of the variables being considered to adult ranges. This restriction occurs whenever heights (and average *f*0/FFs) are considered only for speakers of a given class (i.e., adult females) and not over their entire possible range. All other things being equal, restricting the range of

* Tel.: +1 530 754 0995.
E-mail address: sbarreda@ucdavis.edu

http://dx.doi.org/10.1016/j.wocn.2015.11.004
0095-4470/

variables is very likely to decrease the correlation between them by making the residual error appear larger relative to the amount of systematic variability remaining in the variables after the range restriction (Myers, Well, & Lorch, 2010, p. 453). For example, consider a linear function predicting speaker height from speech acoustics, and imagine that this prediction is accurate to within ± 4 in. When identifying speaker height from a reasonable human range across the entire population (e.g., 3′-6″4′, 42 in. range), 4 in. represents less than 10% of the range, resulting in a reasonably accurate guess. However, if heights in the range considered were restricted to common heights for adult males (e.g., 5′6″ to 6″4′, 10 in. range), 4 in. of error now represents 40% of the range, a substantially larger error for the same underlying process/relationship. This reasoning applies regardless of the underlying estimation error, the full underlying variable ranges, or the restrictions imposed on the covariates by researchers.

The results presented by Pisanski et al. (2014) suggest that although there may be a systematic relationship between size and acoustics even for adult speakers, the magnitude of the systematic component is small relative to the residual error for these restricted ranges. In fact, the degree of systematic variability between height and acoustic parameters in adult speakers may be so small relative to prediction error as to be of limited utility when making any single size prediction. In light of this, it is not surprising that human listeners appear to not be very accurate at identifying the heights of adult listeners from speech. However, although listener judgments tend to be incorrect with respect to veridical speaker sizes, it has frequently been noted that these judgments show remarkable consistency is associating lower *f*0 and FFs with larger speakers, between and within-listeners (Bruckert, Liénard, Lacroix, Kreutzer, & Leboucher, 2006; Collins, 2000; Rendall et al., 2007; Van Dommelen & Moxness, 1995). Essentially, listeners demonstrate sensitivity to the overall covariation of speaker size, *f*0 and the FFs across the entire human population and reliably identify speakers with lower *f*0 and FFs as larger, even though this may lead to incorrect size estimates for adult voices.

Results demonstrating the consistent and predictable use of acoustic cues in making speaker-size judgments indicate that these judgments are the result of the systematic use of the acoustic cues carried by the voices of speakers, on the part of listeners. In light of this, speaker-size judgments (right or wrong) shed light on listener behavior with respect to the acoustic characteristics of voices. In fact, given that noise may overwhelm systematic variability in the relationship between acoustics and speaker size when restricted to an adult range, a focus on accuracy of judgments with respect to the true heights of speakers, rather than on the judgments themselves, could obscure the fact that listeners are behaving systematically with respect to stimulus properties. The remainder of the discussion that follows will focus on the use of spectral cues by listeners in the determination of speaker size. The focus will be on the systematic use of this information by human listeners, and not in any sense on the accuracy of these assessments relative to the real heights of speakers.

1.1. The use of spectral information in the assessment of speaker height

The perception of speaker size has been investigated extensively, using natural, synthetic and resynthesized stimuli, and stimuli ranging in size from isolated vowels, to words and syllables (Barreda & Nearey, 2012; Collins, 2000; Fitch, 1994; Ives, Smith, & Patterson, 2005; Rendall et al., 2007; Smith & Patterson, 2005; Smith, Patterson, Turner, Kawahara, & Irino, 2005; Van Dommelen & Moxness, 1995). These experiments have repeatedly found that the FFs and *f*0 are strongly predictive of perceived speaker size, findings which are mirrored by experiments using statistical classification methods to identify speaker characteristics (Bachorowski & Owren, 1999; Hillenbrand & Clark, 2009). Further, it has been shown that speaker-size judgments can be well modeled by a relatively simple linear combination of measurements of stimulus *f*0 and FF information (Fitch, 1994; Smith & Patterson, 2005). The systematic use of *f*0 information in these judgments is relatively clear: all other things being equal, the speaker with the lower *f*0 is very likely to be identified as taller. However, the use of spectral information (typically indexed using the FFs) may be considerably less straightforward.

In research on the availability or use of size cues in speech, it is common to index variability between speakers using a single parameter meant to represent VTL variation, such as the log-mean FF (Nearey, 1978), mean FF (Pisanski et al., 2014), *F*2 of Schwa (Van Dommelen & Moxness, 1995), formant dispersion/spacing (Collins, 2000), spectral envelope scaling (Smith et al., 2005), or a direct estimate of VTL (Smith & Patterson, 2005). In addition, is it a standard practice to simulate VTL differences between speakers/stimuli by increasing or decreasing stimulus FFs by a single[1] multiplicative scaling-parameter (Assmann, Dembling, & Nearey, 2006; Barreda, 2012; Barreda & Nearey, 2013; Fitch, 1994; Ives et al., 2005; Rendall et al., 2007; Smith et al., 2005; Smith, Walters, & Patterson, 2007).

When investigating size perception, the use of a single parameter (e.g. VTL) to represent the FFs actually present in a stimulus relies on the idea that listeners use the FFs present in a speech sound in order to estimate VTL, and then use this VTL information to make speaker-size judgments. For example, Rendall et al. (2007) suggest that listeners "discriminate size differences based on formant frequency cues to speaker VTL" (1215), Smith et al. (2007) state that "VTL is an important cue to sex and age because it changes with physical size" (3629), Ives et al. (2005) state that "size information in speech is available to the listener and changes in VTL alone [can] produce reliable differences in perceived size" (3822), and Van Dommelen and Moxness (1995) state in their conclusions that "results showed that large VT values, that is low formant frequencies, were interpreted by the listeners as indicating large body dimensions" (283). Although discussion frequently centers on the use of a VTL parameter in size perception, and presents VTL and formant information as roughly equivalent, VTL information is not directly present in the speech signal. This means that VTL information would have to be recovered by listeners on the basis of the FFs, which are directly present in speech sounds.

Research on the perception of speaker size from speech typically involves experimental designs that control for linguistic content or only consider aggregate behavior over a fixed set of tokens. For example, Rendall et al. (2007) and Smith et al. (2005) presented listeners with pairs of stimuli differing in simulated VTL and asked them to identify the taller speaker from the pair. Crucially however, in all cases stimuli were matched for linguistic content within pairs. These designs result in contrasts such as in Fig. 1a, which compares two vowels that differ solely on the basis of simulated VTL differences (i.e., a uniform global shift in all FFs). In these cases, comparing any given formant across the two tokens would yield the same result as using VTL estimates for the two voices: In either case the voice with the lower FFs would likely be identified as taller. This could give the impression that listeners are responding to differences in apparent VTL even if they are simply responding directly to one or more of the FFs present in the stimuli being considered. Although these strategies

[1] Please see the Appendix A for a discussion of the appropriateness of describing and creating stimuli on the basis of a single formant-scaling parameter.

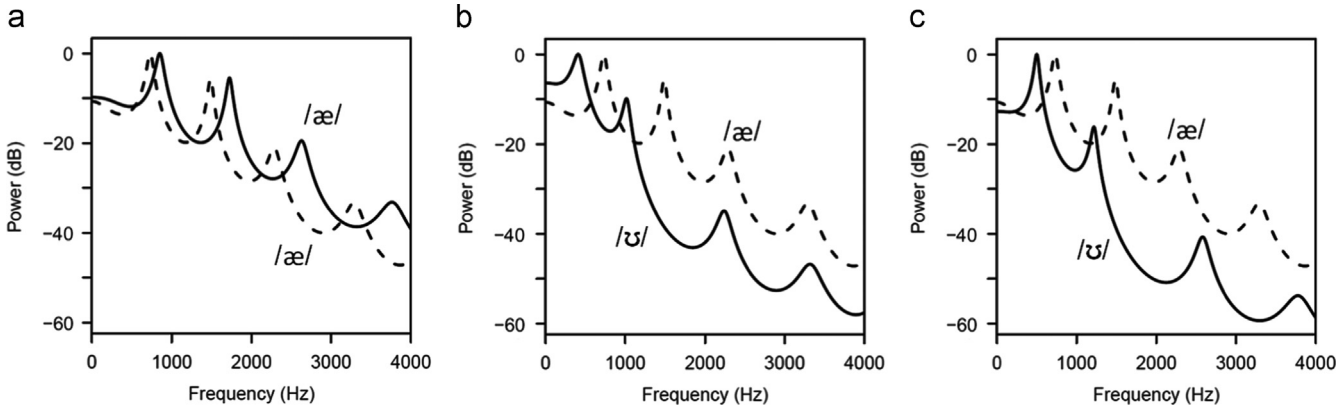

Fig. 1. LPC spectra for synthetic tokens of /æ/ (dashed line) compared with spectra for other vowels at the same or different VTL levels. (a) A comparison with another instance of /æ/ (solid line) whose formants have been uniformly shifted up by 16%.(b) A comparison with an instance of /ʊ/ (solid line) whose formant frequencies are appropriate for the same speaker who produced the /æ/ in the same panel. (c) A comparison with an instance of /ʊ/ (solid line) whose formant frequencies have been shifted up by 16% relative to those of the speaker who produced the /æ/ in the same panel.

might seem interchangeable, they make quite different predictions when one considers size judgments across stimuli with varying linguistic content.

From the perspective of VTL-based size perception, all linguistic tokens produced by a speaker contain the same information, i.e., speaker-dependent VTL information. Further, it is implicit in the very notion of a speaker-dependent VTL estimate that this should be roughly stable for a given speaker, after all, each speaker has a single VTL. Consequently, VTL estimates that show non-trivial variation according to linguistic content would seem to functionally cease being a VTL estimate at all. As a result, as long as roughly the same VTL is implied by two vowels being compared, relative size-judgments should not be significantly affected by vowel category. For example, Fig. 1b presents a case where /æ/ is being compared to /ʊ/, and both vowels have FFs appropriate for a single adult male speaker. Since /æ/ represents the high end of *F*1 and *F*2, while /ʊ/ is closer to the low end of this range, /æ/ will have higher FFs and more high-frequency energy than /ʊ/. However, since both vowels imply roughly the same VTL, listeners should have no basis by which to select one or the other vowel as taller on a consistent basis. Fig. 1c presents another case where /æ/ is being compared to /ʊ/, however in this case the FFs of /ʊ/ have been shifted up by 16% relative to in Fig. 1b (a methodology commonly employed to simulate VTL differences between speakers). If VTL cues are driving the perception of speaker-size judgments, listeners should consistently identify the /æ/ in Fig. 1c as having been produced by the taller speaker (see Fig. 2 for another perspective on these comparisons).

However, as previously noted, VTL cues are not directly present in the speech signal. Instead, even when produced by a single speaker, spectral information can vary a great deal across different linguistic tokens. If listeners use spectral information directly instead of only using it to estimate VTL, then speaker size judgments might be expected to vary according to the spectral characteristics of the stimuli presented. For example, /ʊ/ has much lower *F*1 and *F*2 frequencies than /æ/ within speaker (1b). As a result, large phoneme-specific formant differences might overwhelm the relatively more subtle VTL differences so that the /ʊ/ associated with a shorter VTL can have lower *F*1 and *F*2 than the /æ/ associated with a relatively longer VTL (1c). Because of this, listeners might consistently identify /ʊ/ as taller than /æ/, even when in conflict with VTL information.

As opposed to the comparison in Fig. 1a, those in 1b and 1c allow the following questions to be directly addressed: How does a given formant pattern lead to a given perception of speaker size? Are the FFs used directly so that low FFs are marginally associated with perceived taller speakers? Or rather, do the FFs only affect perceived size by informing VTL estimates, so that a low FF affects perceived speaker-size only to the extent that it suggests a longer VTL? In other words, are FFs used as direct evidence of speaker size, or is a formant-pattern correction employed so that listeners can use something like a VTL scale-estimate for the speaker? As mentioned previously, there is extensive experimental evidence that in situations such as those in Fig. 1a, listeners will tend to identify the vowel with the lower FFs and longer apparent VTL (broken line) as taller.[2] However, the outcome in 1b and 1c is less certain and would provide useful insight into the actual use of spectral information by human listeners when assessing speaker characteristics.

In light of this, investigating the perception of speaker size on the basis of individual stimulus properties, rather than only considering aggregate responses across a range of stimuli, provides a way to test whether listeners are basing speaker height judgments on a VTL parameter, or whether they are only directly responding to stimulus properties. If listeners base size judgments mostly (or entirely) on simulated VTL differences between voices, then that is good evidence that they are, in fact, using a VTL parameter in estimating speaker size. On the other hand, the presence of systematic category-specific preferences or behaviors, in particular if these run contrary to the VTL cues present in the vowels being compared, would suggest that listeners are not basing size judgments only on speaker VTL, and are instead directly responding to the formant-pattern present in the stimulus being considered.

1.2. The higher formants

One factor that could potentially facilitate the use of VTL cues in size perception, thereby minimizing phoneme-specific effects on these judgments should they exist, is a reliance on the higher formants (formants above *F*2). The frequencies of the first two formants are considered to be the primary determinants of perceived vowel quality (Joos, 1948; Miller, 1989; Nearey, 1978; Peterson, 1961). Since these

[2] In the pages that follow, reference will be made to one vowel or the other being identified as taller. Although the apparent speaker that produced the vowel is actually being identified as taller, this leads to a cumbersome phrasing as in, for example “the apparent speaker associated with the voice that produced the /i/ was identified as taller”. Instead, this will simply be stated as “the /i/ was identified as taller”.

Table 1
For each general speaker class (women, men, children), values express the standard deviation between vowel category means for each formant, divided by the average standard deviation of the same formant, within vowel-categories. This value is conceptually similar to an *F*-ratio. Values greater than 1 indicate relatively more between-phoneme variation in the formant, and values less than 1 indicate relatively more between-speaker variation in the formant. Formant values are taken from Hillenbrand et al. (1995), representing formant measurements from 139 men, women and children.

Speaker class	*F*1	*F*2	*F*3	*F*4
Women	3.09	4.49	2.09	0.62
Men	3.22	4.34	2.06	0.54
Children	2.78	4.20	2.03	0.93

formants can vary greatly in order to carry phonetic distinctions, they are not ideally suited for direct usage in speaker-size estimation. On the other hand, the higher formants are largely free of phonetically-motivated variation, they may provide more direct evidence regarding speaker VTL than *F*1 or *F*2 (Deng & O'Shaughnessy, 2003; Rose, 2003; Rose & Clermont, 2001). For example, the potential problems associated with identifying the taller speaker in Fig. 1b–c only manifest when one focuses on the lower formants of the vowel sounds, and the potentially large differences between them across vowel categories. In contrast, if listeners instead focused on the higher formants (*F*3 and *F*4) in Fig. 1, they would be able to respond on the basis of VTL scale information with relative ease.

We may get a sense of the utility of the higher formants in making speaker size judgments by considering variability between and within vowel-categories based on formant number. Hillenbrand et al. (1995) report a large and well-known dataset of formant frequencies (*F*1-*F*4) for 139 men, women and children. In supplemental materials made available online (http://homepages.wmich.edu/~hillenbr/voweldata.html), the authors report FF means for each speaker class (man, woman, child) for each vowel category. The authors also report the amount of within-category (i.e., between-speaker) variation for each formant according to speaker class and vowel category. For a given formant, these values can be used to find the ratio of the standard deviation between different vowel-categories, divided by the standard deviation within vowel categories, presented in Table 1. Since these values express between-category variability divided by within-category variability, they are conceptually similar to the *F*-ratios used in the analysis of variance. Values larger than one in Table 1 indicate that the formant in question varies more between vowel categories than it does between-speakers, within category. On the other hand, values less than one indicate that between-speaker variability (even within a speaker class) is larger than between-phoneme differences for that formant. As seen in Table 1, *F*4 is the only formant to have value lower than 1, indicating that (in this dataset at least) it is the only formant that varies more between-speakers than it does between phonemes.

Although there appear to be general tendencies in the location of *F*4 between vowel categories, these can be overwhelmed by idiosyncratic speaker-specific tendencies. For example, Hillenbrand et al. (1995, Table V) indicate a tendency for back vowels to have lower *F*4 values than front vowels. For example, the average *F*4 for /æ/ is higher than the average *F*4 for /u/ for all speaker classes (men, women and children). However, a look at the individual formant data reveals that 26% of speakers for whom data is available have lower *F*4 values for /u/ than /æ/, and for 38% of speakers these formants are within 100 Hz of each other (about 2.5% of typical *F*4 values). This situation is not limited to this particular comparison: the large amount of between-speaker variability in *F*4 coupled with the small amount of between-category variability means that although there may be general tendencies in the location of *F*4 between vowel phonemes, these tendencies are not deterministically true for all speakers. This is very different in kind to the situation with the lower formants. For example, in most dialects of English we can say with certainty that *F*1 and *F*2 will be higher for /æ/ than for /u/ for a single speaker.

The idiosyncratic placement of the higher formants in human voices, and their lack of a strong effect on vowel quality is reflected in Klatt (1980), where the author outlines the use of what would become known as a Klatt synthesizer. Klatt states that since *F*4 and *F*5 frequencies and bandwidths do not vary much between vowel phonemes within-speaker, they "help to shape the overall spectrum, but otherwise contribute little to intelligibility for vowels" and so they can "be held constant [across phonemes] with little decrement in output sound quality" (980). Effectively, the selection of arbitrary (but plausible) higher formants for use in synthetic vowels mirrors the idiosyncratic locations of these formants in the spectra of real human voices as determined by speaker-specific anatomical and gestural characteristics. In both cases, the frequencies of these formants (*F*4 and *F*5) can be considered to reflect voice-specific characteristics of the real or synthetic voice more so than systematic variability reflective of phonetic content.

The combination of more variability between-speakers than between-phonemes, and the idiosyncratic locations of the higher formants between speakers (in part facilitated by their lack of an important phonetic role) means that formants above *F*3 may be considerably more useful when attempting to estimate speaker characteristics (including VTL) than *F*1 and *F*2. The potential usefulness of the higher formants in determining indexical information regarding the speaker has long been noted in literature on forensic speaker identification (Hayakawa & Itakura, 1995; Greisbach, 1999; Rose, 2003; Rose & Clermont, 2001; Vaňková, 2014). Although the usefulness of the higher formants is well-established when used with pattern-classifiers that attempt to recognize and differentiate speakers, little is known about the distribution, availability, or use of these human formants by human listeners. The scarcity of research into the perceptual use of the higher formants is noted by Donai and Lass (2015) who report on a recent experiment investigating the ability of listeners to identify speaker gender based on isolated vowel sounds that were high-pass filtered at 3.5 kHz. Listeners identified speaker-gender correctly for filtered vowels in 82% of cases, and results were explainable on the basis of higher-formant information.

Although it is expected that listeners will use information related to the first three FFs in vowel sounds, it is not clear if or how they use information in the higher spectrum to make speaker-size judgments. In the event that higher-formant information is used by listeners in estimating speaker-size, it may play a special role in allowing these judgments to be based on VTL, rather than basing these judgments on vowel-specific formant patterns. For example as noted earlier, the potential ambiguity in selecting the larger speaker in Fig. 1 only arises if one focuses on *F*1 and *F*2 instead of on the higher formants. This leads to three general questions regarding the use of the higher formants in arriving at speaker-size judgments: (1) Are formants above *F*3 used by listeners in making speaker-size judgments? (2) If so, does access to these formants increase the ability to estimate VTL accurately, thereby limiting possible phoneme-specific effects in size perception? (3) Is there some minimum number of formants required in order to make systematic use of VTL information in speaker height judgments?

1.3. The current experiment

The objective of the current experiment is to investigate the hypothesis that listeners use something like a VTL estimate in speaker-size judgments, and to investigate the role of the higher formants in making these judgments. The experiment employs a similar experimental task to previous experiments investigating the perception of speaker size (Rendall et al., 2007; Smith et al., 2005). A series of synthetic vowel sounds were created (/i æ ʊ/) and scaled uniformly up or down in frequency in order to simulate VTL differences between voices. These were presented to listeners in pairs, and listeners were asked to indicate which of the vowels sounded like it had been produced by the taller speaker. As opposed to previous experiments however, stimulus pairs were not matched for phonetic content but were allowed to vary within-pair, and results were considered for each trial and not only in aggregate. This resulted in two kinds of trials: same-phoneme trials where the same vowel phoneme was compared at different VTL levels (Fig. 1a), and different-phoneme trials that compared two different vowel phonemes at the same (1b) or different VTL levels (1c). Additionally, to investigate the role of the higher formants, listeners were divided into four groups, each of which was presented with vowels with different numbers of formants (2, 3, 4 and 5-formant vowels).

When listeners are presented with the same phoneme across a pair, they are expected to identify the voice with lower FFs (and longer simulated VTL) as being taller, in accordance with previous experimental findings. If a minimum number of formants (e.g., three) are required in order to assess speaker size, listeners may display disorderly or unpredictable selections in the groups that were presented with vowels with an inadequate number of formants. Listeners exposed to vowels with more formants are expected to be more consistent in their size-judgments, however, this improvement should only continue as long as listeners are using the additional information. In other words, if listeners do not use *F*4 or *F*5 to assess speaker size, there should not be any difference in the results between the groups presented with three-formant vowels and those presented with vowels with more than three formants. On the other hand, credible differences between these groups would present direct evidence that the additional formants are being used by listeners.

When listeners are asked to compare different phonemes, one of two general patterns of results may occur. If listeners rely only on VTL estimates in making speaker-height judgments, these judgments should be explainable solely on the basis of the simulated VTL difference between the voices being compared, and there should not be significant biases towards identifying some vowels as taller than others. However, if listeners determine speaker-height using some alternative strategy, including the consideration of FF information directly and not only to estimate VTL, then results may depend on the specific characteristics of the stimuli being compared in each trial, and not simply on the VTL differences between the voices. Although it is reasonable to expect some variation in VTL and size estimates across vowel categories, if this variability is so large as to meaningfully and consistently affect speaker-size judgments, then it would seem that listeners are not in fact using VTL to estimate speaker size in any real sense.

The precise effect of the higher formants in different-phoneme trials will depend on the presence or absence of a vowel-category effect on judgments, however, it is generally expected that listeners will respond more strongly to VTL cues when more formants are present, since these should facilitate the estimation of speaker VTL. However, just as in same-phoneme trials, behavior should only be affected by the presence of additional formants (e.g., *F*5) if listeners are actually using the formant in question.

2. Materials and methods

2.1. Participants

Participants were 60 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. Participants were divided into four formant-groups based on whether stimuli were 2, 3, 4 or 5-formant vowels. Listeners were randomly assigned to formant groups, resulting in 15 listeners within each group.

2.2. Stimuli

Experiments investigating the perception of speaker size often involve the use of a fixed set of stimuli, scaled up or down uniformly in frequency in order to simulate VTL differences between speakers (Ives et al., 2005; Rendall et al., 2007; Smith et al., 2005). Since formant frequencies (FFs) are directly specified when using synthetic vowels, VTL differences for synthetic vowels may be simulated by simply multiplying the specified FFs by a fixed constant (Barreda & Nearey, 2013; Fitch, 1994; Nearey, 1989). The second approach was adopted here, and VTL differences between speakers were simulated by increasing/decreasing the FFs used to create synthetic vowel sounds by a single scale factor.

Stimuli consisted of the vowels /i æ ʊ/, based on average productions of adult male speakers of Edmonton English (Table 2). These baseline vowels can be thought of as representing the voice of a single synthetic 'speaker', with a single VTL and complete with an idiosyncratic placement of the higher formants, stable for the speaker. Differences in VTL were simulated by increasing/decreasing the values in Table 2 in equal logarithmic steps roughly equal to differences of 8% (0. 077 log-Hz) between adjacent VTL levels. The baseline FF

Table 2
Formant frequencies for vowels representing the central VTL step for the low *f*0 voices (L2 in Fig. 2a).

	/i/	/æ/	/ʊ/
***F*1**	295	755	458
***F*2**	2262	1576	1100
***F*3**	2900	2441	2392
***F*4**	3500	3500	3500
***F*5**	4500	4500	4500

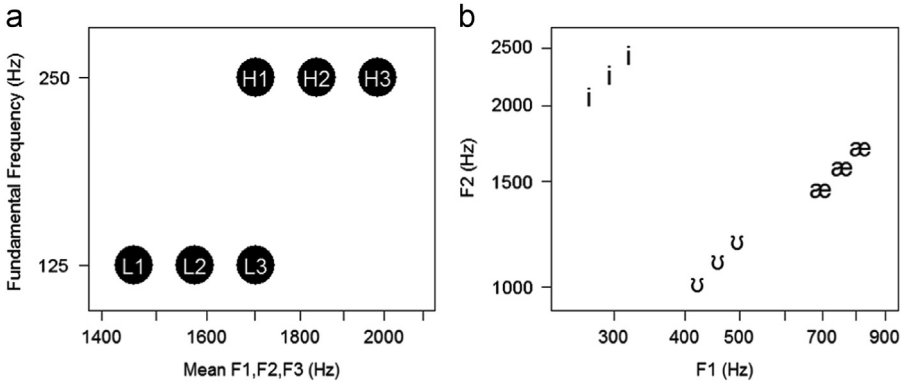

Fig. 2. (a) Locations of the low *f*0 (L) and high *f*0 (H) stimulus voices in an *f*0 × average FF space. Numbers indicate VTL step within the *f*0 level. The experimental task asked listeners to compare voices that differ solely on the basis of FF/VTL differences (horizontal differences) and never *f*0 differences (vertical differences). (b) The formant-space locations of the 9 low-*f*0 vowels. In general, within-categories, proximity to the lower left-hand corner indicates a larger speaker. However, between-category variation is such that a global consideration of proximity to this corner in determining speaker size could result in /ʊ/ being identified as larger than /æ/, regardless of the VTLs implied by the vowels.

values in Table 2 were shifted down one step, and shifted up 3 steps, resulting in a 5-step VTL continuum (of which the baseline voice is the second step).

One issue that arises when simulating a wide range of VTL differences, especially when *f*0 is fixed across a wide range of VTLs, is that this may result in inappropriate *f*0 x VTL combinations (e.g., the FFs of a small child with the *f*0 of an adult male). As a result, it would not be possible to use a single *f*0 level for the entire VTL continuum that would result in roughly appropriate *f*0 levels for all voices. To counteract this problem, the first three VTL steps were combined with an *f*0 that decreased linearly from the beginning to the end of the vowel from 130 Hz to 120 Hz, and the last three VTL steps were synthesized with *f*0s that decreased linearly from 260 Hz to 240 Hz. This resulted in 6 unique combinations of VTL scaling and *f*0 level, each represented by 3 vowels (18 unique vowel stimuli). The locations of these voices on an *f*0 × VTL space are presented in Fig. 2a. Fig. 2b plots the formant-space locations of the low *f*0 vowels in a manner that contrasts between-category variability in FFs with variability due to simulated differences in VTL (within-category variability).

The simulated VTL differences resulted in a range of approximately 16% in FF differences between the highest and lowest VTL steps within each *f*0 level. This difference is large but appropriate based on ranges reported for adult male voices. For example, the log-mean formant frequency for the first four formants for every instance of /i/ in adult male speakers in Hillenbrand et al. (1995) shows a range of 18% and the same value for Peterson and Barney (1952), for the first three formants, shows a range of 22%. Furthermore, the speaker with the lowest FFs in the Hillenbrand et al. (1995) data produced /i/ with an *f*0 of 142 Hz, while the speaker with the highest FFs for /i/ produced it with an *f*0 of 133 Hz, indicating that a range of roughly 16% in FFs with a stable *f*0 is appropriate given the distribution of these characteristics in real voices.

The vowel categories /i æ ʊ/ were chosen for two reasons. First, they represent a broad range of *F*1 and *F*2 values and so allow for tests comparing the effects of relatively subtle VTL differences with those of relatively large differences in *F*1 and *F*2 between categories. Second, because of their locations in the vowel space of the Edmonton English dialect, these vowels are particularly resistant to vowel-category changes resulting from shifts in *f*0 and/or simulated VTL. After synthesis and prior to experimentation, speakers of the local dialect confirmed the lack of vowel-category shifts in the resulting stimuli.

2.2.1. Synthesis details

Vowels were synthesized using a Klatt-style (Klatt, 1980) parametric synthesis program implemented in MATLAB. *F*4 and *F*5 were fixed across vowel categories at values roughly appropriate for an adult male speaker, following conventions when using synthetic vowels (Klatt, 1980; Nearey, 1989). Each consecutive formant above *F*5 was set to 1000 Hz higher than the previous one, up to the 11th formant (Holmes, 1983). Although no vowel in the experiment contained more than 5 formants, the inclusion of these formants is necessary when synthesizing vowels at a high sampling frequency to prevent differences in spectral slope that can arise from changes in the distance between the highest specified formant and the Nyquist frequency (Holmes, 1983). Since the highest specified formant varied as a function of the VTL scaling applied to the formant pattern, the sampling frequency used for synthesis was varied so that the Nyquist frequency would fall halfway between the 11th formant, and the expected frequency of the 12th formant, had it been specified. 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. Vowels were presented to listeners in pairs separated by 250 ms of silence.

2.2.2. Number of formants

Each vowel to be used in the experiment was synthesized a single time for a single combination of VTL and *f*0 levels, resulting in 18 unique, eleven-formant vowels (6 VTL and *f*0 combinations × 3 vowel categories). The vowels to be used for the different formant-groups were all based on these original vowels. The procedure for the creation of vowels for each formant-group was as follows. First, the vowel sound was low-pass filtered with a cut-off frequency set at halfway between the highest formant to be included, and the first formant to be excluded (see Fig. 3). After filtering, each vowel was resampled to a sampling frequency of 22,050 Hz. As mentioned previously, vowels were originally all synthesized with 11 formants simply to ensure that there were not large differences in spectral tilt associated with the different number of formants in the different groups, or with the different synthesis sampling frequencies used for the various VTL scaling levels. The result of these manipulations is that the vowels presented to listeners in the different formant groups differed only in the presence/absence of a given number of additional formants.

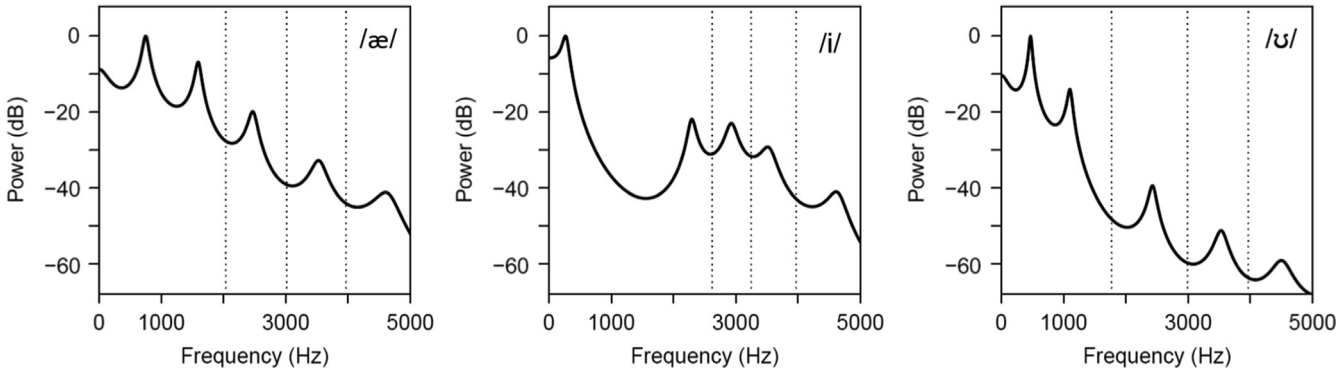

Fig. 3. LPC spectra of stimulus vowels for the baseline, low *f*0 voice, L2 in Fig. 2b. Formant frequencies for these vowels are provided in Table 2. Dotted lines show filter cutoff frequencies used for the creation of 2, 3, 4 and 5 formant vowel sounds.

2.3. Procedure

Listeners were presented with vowel sounds in pairs and asked to indicate which of the vowels sounded like it had been produced by a taller speaker. Listeners were only asked to compare vowel sounds synthesized at the same *f*0 level, and never across *f*0 levels. For example, this means that in Fig. 2a, low *f*0 voices (L) would be compared to other low *f*0 voices, but never to voices at the higher *f*0 step (H). Consequently, relative speaker-size judgments provided by listeners can be said to result from spectral differences (simulated VTL or vowel category), and not from information related to *f*0.

Each stimulus vowel was combined with every other sound at its *f*0 level (except for itself), resulting in 72 combinations for each *f*0 level. Trials were blocked by *f*0 level so that listeners first responded to vowels at one *f*0 level, and then responded to vowels at the other *f*0 level. These blocks were separated by a self-timed pause. Block order was balanced across subjects. Each combination of vowel pair and VTL difference was repeated 3 times, resulting in 216 responses per block, and 432 overall. Within each *f*0-level block, vowel pairs were presented randomized along all stimulus dimensions, blocked by repetition. Vowels were balanced for order across all stimulus dimensions and repeated across voice-size groups resulting in 12 responses per listener for each combination of vowel pair and VTL difference (3 repetitions $\times$ 2*f*0 levels $\times$ 2 orders).

Listeners were told that they would be hearing a series of synthetic voices patterned after male voices from a wide range of ages, from children to adults. This was done to minimize the possibility of an effect for perceived speaker gender, in particular for the high-*f*0 voices. Listeners were instructed that they would be hearing the vowels in the words ‘heed’, ‘had’ and ‘hood’ presented in pairs, and that they had to decide which of the two vowels had been produced by a taller speaker. Sounds were presented over headphones, in a sound-attenuated booth. Listeners provided responses by clicking on a specially designed graphical user interface that contained two response buttons: the button on the left labeled ‘First Voice’, or the one on the right labeled ‘Second Voice’. The response buttons were labeled ‘HEED’, ‘HOOD’ or ‘HAD’ as appropriate given the trial to further remind listeners of which response button corresponded to which vowel. The user interface also contained a button marked ‘Replay’ that allowed listeners to replay the presented stimulus up to three additional times for each trial. After the listener had made a selection, the next stimulus was played after a one second pause. Every listener completed both blocks of the experiment.

2.4. Statistical analysis: Bayesian multilevel logistic regression

The experimental task asked listeners to decide whether the first or second voice in each pair sounds taller, resulting in a dichotomous outcome variable. Results were analyzed using a Bayesian multilevel logistic regression model, which simultaneously models the results of individual subjects, while also pooling information across all subjects to estimate group-level effects. Analyzing data using a Bayesian multilevel model offers several advantages relative to more traditional maximum-likelihood estimation techniques, such as detailed information about credible values for all parameters, jointly credible values (or differences in values) for groups of parameters, and protection against multiple comparisons without the resizing of credible/confidence intervals (Gelman, Hill, & Yajima, 2012; Kruschke, 2014, 2010; Kruschke, Aguinis, & Joo, 2012).

Bayesian inference relies on the posterior distribution of parameter values given the data and the prior probabilities of the parameters. Although the exact form of the posterior cannot usually be determined analytically for even moderately complex multilevel models, these distributions may be approximated using Markov Chain Monte Carlo (MCMC) methods as implements by software such as JAGS (Plummer et al., 2003). These methods find the distribution of jointly-credible values for all model parameters by taking a series of random ‘steps’ through the joint parameter space. After a given number of these steps, the result is a ‘chain’ of parameter values which can be used to assess credible, and jointly-credible, values for parameters or combinations of parameters. If a value of interest (e.g. 0) is very atypical given the distribution of values in the posterior distribution (the ‘chain’) it is deemed to not be a credible value for that parameter. Conversely, if the value of interest is within the bounds of the values in the chain, it is deemed to 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).

Results will be considered in two general cases: same-phoneme and different-phoneme trials. In same-phoneme trials, each pair consisted of the same vowel phoneme presented at different VTL levels. These two kinds of trials will be analyzed independently. In each case, trials will only be considered as a function of VTL and vowel-category differences between voices, meaning results will be collapsed across *f*0 levels.

2.4.1. Same-phoneme trials

Same-phoneme trials featured clear expectations regarding listener behavior. Based on previous experiments, it is expected that listeners would identify the speaker with the longer simulated VTL (i.e., the lower average FFs) as being taller. Since this expectation is supported by a good deal of experimental evidence, these responses will be considered 'successful' in the sense that the listener is conforming to expectations. The adoption of the term 'successful' rather than 'correct' for trials where listeners associated a longer simulated VTL with a taller speaker is a conscious decision to avoid the implication that these trials represent the accurate use of acoustic cues with respect to some real speaker height. The probability of a successful response in any given trial can be modeled as a logistic function of the parameter θ as in (1).

The θ parameter can be broken down further using an ANOVA-style decomposition as in (2). This models the parameter θ as a linear combination of the intercept (α_0) and groups of deflection terms indicating deviations from the overall mean (α_0) due to the main effects and interaction terms, with a group of deflection terms associated with each predictor and interaction term. The predictors included in the model are: formant group (α_F, 4 levels: 2, 3, 4, 5), vowel category (α_V, 3 levels: /æ i ʊ/), absolute VTL difference ($\alpha_{\Delta VTL}$, 2 levels: 1 step or 2 steps), subject (α_S, 60 levels), the formant group by vowel category interaction ($\alpha_{F\times V}$, 12 levels), the VTL difference by formant-group interaction ($\alpha_{\Delta VTL\times F}$, 8 levels), and the VTL difference by vowel interaction ($\alpha_{\Delta VTL\times V}$, 6 levels).

$$p(\text{success}_i) = \text{logistic}(\theta_i) = \frac{\exp(\theta_i)}{\exp(\theta_i)+1} \tag{1}$$

$$\theta_i = \alpha_0 + \alpha_F + \alpha_V + \alpha_{\Delta VTL} + \alpha_S + \alpha_{F\times V} + \alpha_{\Delta VTL\times F} + \alpha_{\Delta VTL\times V} \tag{2}$$

In order to make the parameter estimates in a decomposition as in (2) identifiable, the deflections associated with main effects in (2) were constrained to sum to zero around α_0, and the interactions were constrained to sum to zero within-factors, as described in Kruschke (2014, Chapter 20). As recommended in Gelman and Hill (2006, p. 494) each group of deflection terms associated with the parameters in (2) was modeled as coming from a separate, higher-level distribution with a mean of zero (following from the sum-to zero constraints imposed on the parameters) and a parameter-specific variance also estimated from the data. For example, the 60 subject deflections (α_S) were modeled as coming from a normal distribution with a mean of μ_S (set to 0) and a variance of σ_S^2, while the 12 formant-group by vowel interaction terms ($\alpha_{F\times V}$) were modeled as coming from a normal distribution with a mean of $\mu_{F\times V}$ (set to 0) and a variance of $\sigma_{F\times V}^2$. Vague priors were used in all cases: uniform priors were used for all the higher-population variance parameters (Gelman, 2006), and the intercept term (α_0) was given a prior mean of zero and a prior variance of 16, which is quite large relative to typical variation in logistic coefficients.

In all, 103 parameters (including 7 variance parameters, one for each bundle of deflection coefficients in Eq. (2)) were estimated from the data. The posterior samples for all parameters were generated using JAGS (Plummer et al., 2003) and R (R Core Team, 2015). Four independent chains were run, with each chain being a total of 2500 steps in length, for a total of 10,000 steps. A 10,000 step burn-in period was used and chains were thinned every 200th step to reduce autocorrelation in the chains and to maintain a reasonable file size (500,000 total steps were run). The chains mixed well, with the effective sample sizes of all lower-level parameters (controlling for autocorrelation in estimates) being nearly 10,000.

2.4.1.1. Predictions regarding directions of effects. Success in same-phoneme trials is expected to differ primarily according to the VTL difference between voices ($\alpha_{\Delta VTL}$), with larger VTL differences leading to more successful responses. So for example, success rates should be higher for $\alpha_{\Delta VTL=2}$ than for $\alpha_{\Delta VTL=1}$, which also means that $\alpha_{\Delta VTL=2} - \alpha_{\Delta VTL=1}$ should have a positive and credibly non-zero value.

Of particular interest in same-phoneme trials is the change in success rates according to the number of formants present in vowel stimuli (α_F). As outlined in Section 1.2, a difference in performance based on the presence of a given higher formant (e.g. *F*5) presents evidence that listeners use that formant in assessing speaker size. On the other hand, a lack of difference in performance between formant-groups differing by a single formant would suggest that the formant differentiating the groups is not being used by listeners. This may be assessed by investigating the difference in deflection-parameter estimates between any two given formant groups. For example, a credible difference in success rates based on the presence of *F*4 may be assessed by inspecting the distribution of the difference between the deflection estimates for the 3 and 4 formant groups, $\alpha_{F=4} - \alpha_{F=3}$, while $\alpha_{F=5} - \alpha_{F=4}$ would yield information regarding the use of *F*5 if present.

2.4.2. Different-phoneme trials

In different-phoneme trials (i.e., trials where the pair contrasted different phonemes) there was not a clear expectation of how listeners would behave. To explore these trials with as few assumptions as possible, /æ/ was selected as the reference phoneme in these trials, and /ʊ/ and /i/ acted as the alternative vowels. This means that the different-phoneme analysis focused on the trials featuring /æ/-/i/ and /æ/-/ʊ/ comparisons. The VTL difference between vowels was coded as a continuous predictor expressing the difference in VTL step between the reference vowel and the alternative vowel ($VTL_{Reference} - VTL_{Alternative}$), which could take on integer values between -2 and 2. This coding leads to a logistic function as in (3), where the probability that the reference vowel (/æ/) was identified as taller is broken down in (4) into VTL difference effects (β) and an intercept term (α).

$$p(\textit{reference vowel is taller}_i) = \text{logistic}(\theta_i) = \frac{\exp(\theta_i)}{\exp(\theta_i)+1} \tag{3}$$

$$\theta_i = \alpha_i + \beta_i \times VTL_i \tag{4}$$

Since positive VTL difference values indicate that the reference vowel had a longer VTL, VTL difference is expected to be positively related to the selection of the reference vowel as taller. The intercept term reflects the probability that the reference category will be selected as taller when the VTL difference between voices is equal to zero, indicating that vowels have the same VTL. However, because the only linear predictor included in the model (VTL difference) is centered at zero, a non-zero intercept term will also reflect an overall tendency to identify one vowel as taller than the other in a given pair. As a result, the intercept terms are indistinguishable from response

biases, and variability in estimated intercept terms can be interpreted as variability in these biases. The slope and bias/intercept terms in (4) can be further decomposed as in (5) and (6).

$$\alpha_i = \alpha_0 + \alpha_F + \alpha_V + \alpha_S + \alpha_{F\times V} \tag{5}$$

$$\beta_i = \beta_0 + \beta_F + \beta_V + \beta_S + \beta_{F\times V} \tag{6}$$

Eq. (5) models vowel biases as linear combinations of the mean intercept (α_0) and predictors: formant group, (α_F, 4 levels: 2, 3, 4, 5), the alternative vowel category (α_V, 2 levels: /i ʊ/), subject (α_S, 60 levels), and the formant group by alternative vowel interaction ($\alpha_{F\times V}$, 8 levels). The analogous decomposition holds for (6). Just as in the same-phoneme trials, each of the deflection terms associated with the effects in (5) and (6) was constrained to sum to zero around their respective overall means, and interactions were constrained to sum to zero within-factors. For example, the four α_F terms were constrained to sum to zero about α_0, while the four β_F terms were constrained to sum to zero around β_0. As with the same-phoneme analysis, each group of deflection terms in (5) and (6) was modeled as coming from a separate, higher-level distribution with a mean of zero and a parameter-specific variance, which were also estimated from the data. Uniform priors were used for all the higher-population variance parameters, and the intercept terms (α_0, β_0) were given a prior means of zero and a prior variances of 16.

In all, 158 parameters (including 8 variance parameters, one for each group of deflection coefficients) were estimated from the data. Posterior samples for all parameters were again generated using JAGS and R. The same approach was used as for the same-phoneme model: Four independent chains were run, with each chain being a total of 2500 steps in length, for a total of 10,000 steps. A 10,000 step burn-in period was used and chains were thinned every 200th step to reduce autocorrelation in the chains and to maintain a reasonable file size (500,000 total steps were run). The chains mixed well, with the effective sample sizes of all lower-level parameters (controlling for autocorrelation in estimates) being nearly 10,000.

2.4.2.1. Predictions regarding directions of effects. The effects in different-phoneme trials that most directly address the research questions posed here are the vowel-category biases (α_V), the interaction between these biases and the number of formants presented ($\alpha_{F\times V}$), and the effect of number of formants on the use of VTL differences (β_F). If listeners only use VTL cues in making speaker-size judgments, the intercept term should not be affected by vowel category so that all the α_V terms should only randomly vary around zero. In other words, if listeners base their size estimates only on VTL cues, phoneme-specific biases in size perception should not exist. As for the $\alpha_{F\times V}$ terms, if the higher formants facilitate the use of VTL cues, these may result in a diminishing of any phoneme-specific biases, should they be present.

In terms of the VTL difference effects (β), a positive VTL difference indicates the reference vowel has a longer simulated VTL ($VTL_{reference} - VTL_{Alternative}$) so that a positive VTL difference should result in the reference vowel being perceived as taller at a higher rate. This means slopes should be positive if listeners are behaving roughly as expected. If more formants in a stimulus facilitate size estimation by providing more information, the VTL difference slope will be higher for groups exposed to vowels with more formants. However, as noted previously, these differences should only manifest if listeners actually respond to the formant differentiating the groups being compared. Credible differences in these slopes between groups may be assessed by inspecting differences in β_F terms. For example, a credibly non-zero value for $\beta_{F=4} - \beta_{F=3}$ would indicate that listeners use *F4* in estimating speaker size since behavior is credibly affected by the presence of the formant.

3. Results

Results will be considered in two general cases: same-phoneme and different-phoneme trials. In same-phoneme trials, each pair consisted of the same vowel-phoneme presented at different VTL levels. In different-phoneme trials, each pair consisted of two different phonemes, presented at the same or different VTL levels.

Inferences regarding different effects on the perception of speaker size will be made with the use of the models outlined in Section 2.4. The posterior distributions of parameters will be summarized using the upper and lower bounds of the 95% HDI (Highest Density Interval), which indicates the range of 95% of the posterior distribution of a parameter such that every point inside the range is more probable than every point outside the range (Kruschke, 2010). The mean and the percent of samples above or below zero (as appropriate) will also be reported. Inferences regarding differences between parameters will be assessed by finding the difference between parameter estimates at each step in the chain, and by inspecting the distribution of these differences just as with the distribution of the original parameters.

Before continuing, some pertinent characteristics of logistic regression coefficients will be discussed. Logistic regression coefficients are expressed in logits, which are the log-odds, $\ln\left(\frac{p}{1-p}\right)$, that an event p will occur. A logit value of 0 equals a probability of 0.5 and probabilities approach 1 as logit values approaches positive infinity, and 0 as values approach negative infinity. Probabilities may be found from logits by replacing the θ term in Eq. (1) (or 3) with the appropriate logit value. Logit values expressing mean performance for a particular group/condition have a straightforward interpretation as probabilities via the use of Eq. (1). However, care must be taken when considering changes in logits since equal, linear differences in probabilities may be dramatically different when considered in logits, depending on their location between 0 and 1. For example, an increase in performance from 70% to 71% is equal to an increase of 0.048 logits, while an increase from 90% to 91% equals 0.116 logits, and an increase from 95% to 96% is equal to 0.234 logits. For this reason, logit values will also be presented as probabilities in figures where these describe fixed probabilities for specific listening situations. However, these will not be provided when they express differences between groups, since these cannot be interpreted as changes in probability in the absence of some specific baseline value.

This also highlights another important aspect of working with probabilities, namely that relatively small differences in probability near the bounds of the range (0 and 1) can in fact be much more meaningful than seemingly larger differences in probability towards the center of the range (0.5). As a result, summarizing differences in performance on the basis of the arithmetic mean of probabilities may obscure important, yet potentially subtle, differences in listener behavior, and a focus on the logistic analysis is more appropriate.

3.1. Same-phoneme trials

All subjects responded to 108 same-phoneme trials, divided equally among the three vowel categories, for a total of 6480 responses. Same-phoneme trials represent the sorts of comparisons typically employed in previous experiments investigating the perception of speaker height. These trials serve two primary purposes: to confirm that listeners behave as expected in same-phoneme trials and associate longer simulated VTLs with larger speakers, and to investigate the use of the higher formants in the determination of speaker size. As mentioned in Section 2.4.1, trials where the listener identified the vowel with the longer simulated VTL as taller will be referred to as successful trials in that listeners are successfully using simulated VTL cues in the expected manner. This terminology is being adopted in order to avoid any implication that a real height is being correctly estimated.

Average success rates for each listener are presented across different experimental factors in Fig. 4. Please note that in the figure, each listener contributes one point to each vowel category and each VTL difference within formant-group, but that different listeners are represented in each formant group. Listeners exhibited a strong tendency to identify the voice with the longer simulated VTL as taller, doing so in 79.7% of all same-phoneme trials.[3] In light of these results, and in keeping with the earlier discussion, same-phoneme trials where the voice with the lower VTL was identified as being taller will be considered successful responses.

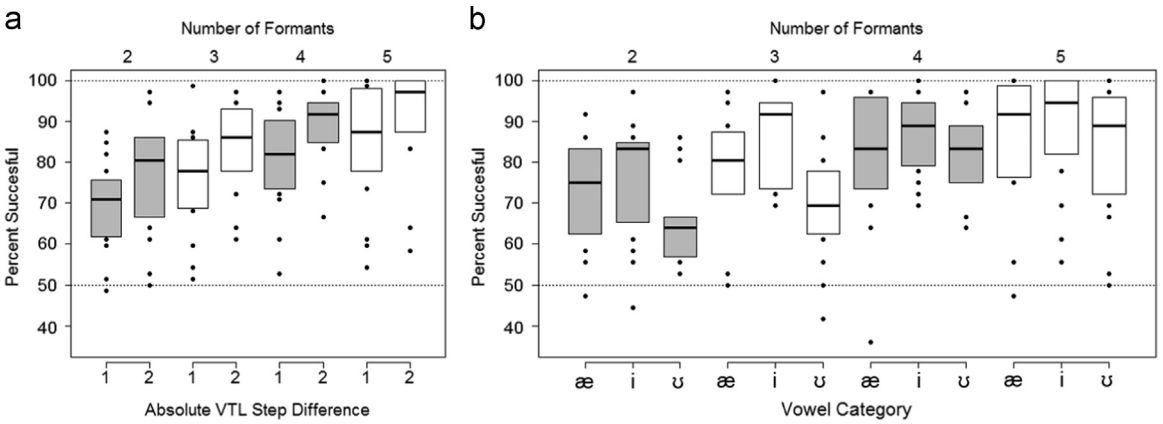

Fig. 4. Points represent the percent of successful responses averaged within-participant, presented according to different combinations of experimental factors. (a) Success rates for different VTL differences across formant groups. (b) Differences in success rates between vowel phonemes across formant groups. Boxplots indicate medians and interquartile ranges; whiskers have been omitted so that outliers can be clearly seen.

Fig. 4 suggests that success rates vary systematically according to formant group, the absolute VTL difference between the voices being compared, and vowel category. Analysis of same-phoneme trials will be based on posterior samples of parameter values using the model outlined in Section 2.4.2. First, data will be analyzed with a Bayesian Analysis of Variance using the finite-population standard deviations as suggested in Gelman (2005). The finite-population standard deviations are estimated for each group of predictors by finding the standard deviation for the relevant predictors at each step in the chain, and then inspecting the distribution of these estimates across all of the steps in the chain. For example, the variance of the formant group parameters can be estimated by finding the variance of the $\alpha_{F=2}$, $\alpha_{F=3}$, $\alpha_{F=4}$, and $\alpha_{F=5}$ parameters at each step of the chain (3 degrees of freedom). This approach focuses on finding the primary sources of variation in success rates, and on estimating their relative magnitudes, rather than focusing on accepting or rejecting potential sources of variance as zero or non-zero (Gelman, 2005). As recommended in Gelman and Hill (2006, Chapter 22) this analysis will then be used to guide the inspection of effects and contrasts as appropriate.

The results of this analysis are presented in Fig. 5. Results indicate that success rates (i.e., the association of longer simulated VTLs with taller speakers) varied primarily as a function of individual differences between participants (α_S), the number of formants in the vowel (α_F), the VTL difference between the vowels ($\alpha_{\Delta VTL}$), and the vowel category presented (α_V). The largest source of variance is differences in the individual abilities of different listeners which, as is evident in Fig. 4, is quite substantial. For example, within the 5-formant vowel group of listeners, six listeners responded successfully in at least 105 of 108 same-phoneme trials, and another three listeners responded successfully in fewer than 68 trials. There is no evidence that any of the interactions considered reflect important sources of variation in success rates as the 95% HDIs for all interaction terms contain values very near to zero. This suggests roughly additive effects for vowel, formants and VTL differences on success rates. Fig. 5b presents HDIs for posterior distributions of estimates for selected parameters based on the results presented in Fig. 5a.

As discussed in Section 2.4.1.1, it was expected that success would vary as a function of the VTL difference between voices. This effect may be assessed by subtracting the difference between the estimates for trials with different VTL differences, $\alpha_{\Delta VTL=2} - \alpha_{\Delta VTL=1}$, which has a mean difference of 0.563 with none of the differences being less than zero (HDI = 0.419, 0.718; 0% < 0), confirming this expectation. Unexpectedly, success also varied as a function of vowel category with the taller voice being identified most successfully for /i/, followed by /æ/ and then /ʊ/. Both the difference in successes between /i/ and /æ/ ($\alpha_{V=i} - \alpha_{V=æ}$, mean = 0.352; HDI = 0.172, 0.532; 99.97% > 0) and the difference in successes between /æ/ and /ʊ/ ($\alpha_{V=æ} - \alpha_{V=ʊ}$, mean = 0.309; HDI = 0.147, 0.481; 100% > 0) were credibly greater than zero.

Of primary interest in same-phoneme trials were differences in success rates according to formant group, and in particular, differences in these rates associated with the presence or absence of higher formants. One question was whether listeners could make consistent

[3] The reason for the apparent disparity between this value and the values presented in Fig. 5b is that calculating the average proportion of successful trials using the arithmetic mean underestimates performance by undervaluing values near 100%. Please see the discussion at the beginning of Section 3.

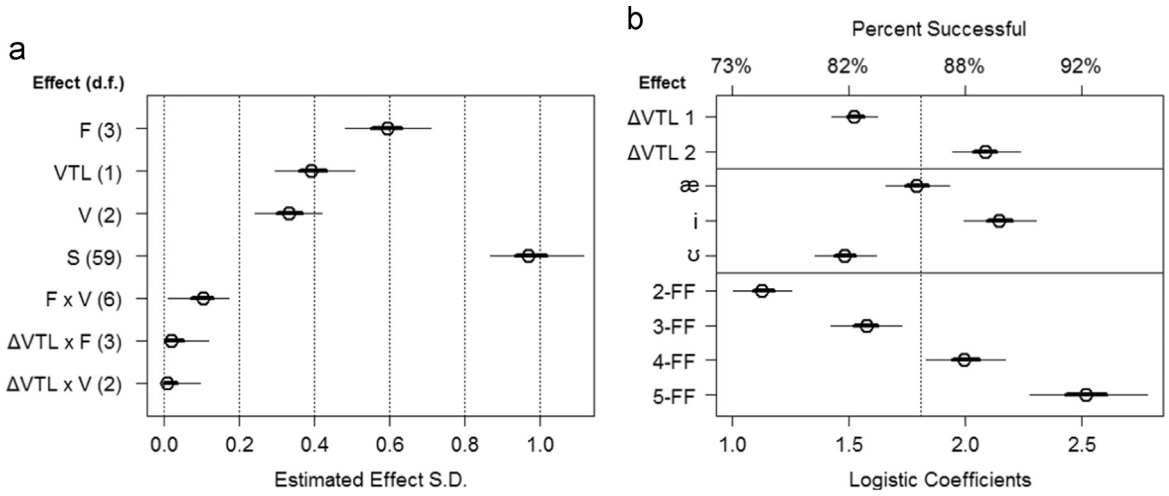

Fig. 5. (a) Results of a Bayesian Analysis of Variance using the finite-sample standard deviations for each effect. Effects are indicated using the appropriate subscripts. (b) Posterior distributions of parameter estimates for selected predictors. These values represent the sum of the appropriate sum-to-zero deflection coefficient and the α_0 term in Eq. (2), indicated with the vertical dotted line. In both panels circles indicate means, bold lines indicate 50% HDI, and thin lines indicate 95% HDI for posterior distributions.

speaker-size judgments in the absence of higher formants (i.e., the 2-formant vowel group). Results indicate that performance for this group was well above chance, with estimated success rates of 75.6% ($\alpha_{a0}+\alpha_{F=2}$, mean=1.13; HDI=1.00, 1.26; 100% > 0). Another important question was whether or not listeners use *F*4 and *F*5 when estimating speaker size. Since the formant-group stimuli differed only in the presence/absence of higher formants, credibly different success rates between formant groups offer strong evidence that listeners use the additional formant in making speaker size judgments. Results indicate credibly non-zero differences in success rates between 2 and 3 formant groups ($\alpha_{F=3}-\alpha_{F=2}$, mean=0.450; HDI=0.253, 0.639; 100% > 0), 3 and 4 formant groups ($\alpha_{F=4}-\alpha_{F=3}$, mean=0.419; HDI=0.191, 0.640; 99.99% > 0) and 4 and 5 formant groups ($\alpha_{F=5}-\alpha_{F=4}$, mean=0.525; HDI=0.244, 0.838; 99.99% > 0). In each case, the presence of an additional formant results in a higher rate of associations between lower FFs and larger speakers. These results indicate that listeners will use formants above *F*3, if available, in order to make speaker size judgments that conform to the expected association between a longer simulated VTL and a perceived larger speaker.

3.2. *Different-phoneme trials*

All subjects responded to 216 different-phoneme trials for the /i/-/æ/ and /ʊ/-/æ/ vowel pairs, for a total of 12,960 responses across all listeners for these pairs. Unlike in same-phoneme trials, there was not a clear expectation for how listeners would behave. On the one hand, if listeners base speaker size judgments solely on VTL cues, the vowel with the longer simulated VTL should be identified as taller with no systematic preferences for one vowel category over another. On the other hand, if listeners consistently identify some vowels as taller than others, especially when in conflict with apparent VTL differences, this would present evidence that speaker-size judgments are not solely based on VTL cues. To investigate the presence of vowel-category effects on speaker-size judgments, different-phoneme trials were considered in terms of whether the reference vowel /æ/ was chosen as taller. Within-participant averages are presented in Fig. 6. Please note that in the figure, each listener contributes one point to each of the vowel pairs at each VTL difference, but that different listeners are represented in each formant group.

Fig. 6 shows that listeners do not have a strong preference for selecting /i/ as taller than /æ/, selecting /i/ in 55.4% of all cases. However, there is a strong tendency to identify /ʊ/ as taller than /æ/, and /ʊ/ was selected in 74% of cases overall. This tendency persisted even when in conflict with relatively large VTL cues. Although in Fig. 6a the selection of the larger speaker varies primarily according to VTL cues, in 6b even large VTL differences barely lead /æ/ to be selected as the taller vowel. For example, even when /ʊ/ had a simulated VTL that was 2 steps smaller than /æ/ (a difference of approximately 16%), listeners still identified /ʊ/ as the taller vowel in 53.5% of cases.

Please note that although the magnitude and direction of biases towards one or another vowel category will depend on the selection of the reference vowel, it is not their exact magnitude or direction but rather their very presence that indicates that these decisions are not made solely using VTL cues. Essentially, if listeners determined speaker size solely on the basis of VTL all results in Fig. 6 should look something like those of the 2 or 3 formant groups in 6a: Variation would be primarily according to VTL and listeners would have no strong preference for one or the other vowel in the absence of a VTL differences between voices. Instead, there are clear, systematic differences between the results presented in 6a and 6b.

The model presented in Section 2.4.2 independently models the effects of VTL differences between vowels (Eq. (6)) and biases towards specific categories (Eq. (5)) as a function of formant group and the alternative vowel. These can be visualized with the aid of Fig. 6. The slope terms express variation between boxes according to VTL differences, but also variability in these slopes between different formant groups and vowel pairs. For example, consider the greatly differing slopes for 3-formant vowels in Fig. 6a and b. The intercept terms express the distance between boxes and the horizontal line at 50% when VTL differences are equal to zero, while also reflecting overall tendencies to select one vowel or the other as taller (i.e., response biases). For example, again consider the greatly differing values for 3-formant vowels at 0 VTL difference in Fig. 6a and b.

Analysis of the different-phoneme trials will be based on posterior samples of parameter values using the model outlined in Section 2.4.2. Data will again be analyzed with a Bayesian Analysis of Variance using the finite-population standard deviations in order to guide a more detailed analysis of the estimated effects. Fig. 7 shows the results of the Bayesian ANOVA, presented in terms of (a) intercept effects and (b) VTL slope effects. Variability in the intercept will be considered first. Fig. 7a indicates a large non-zero average intercept (α_0) that

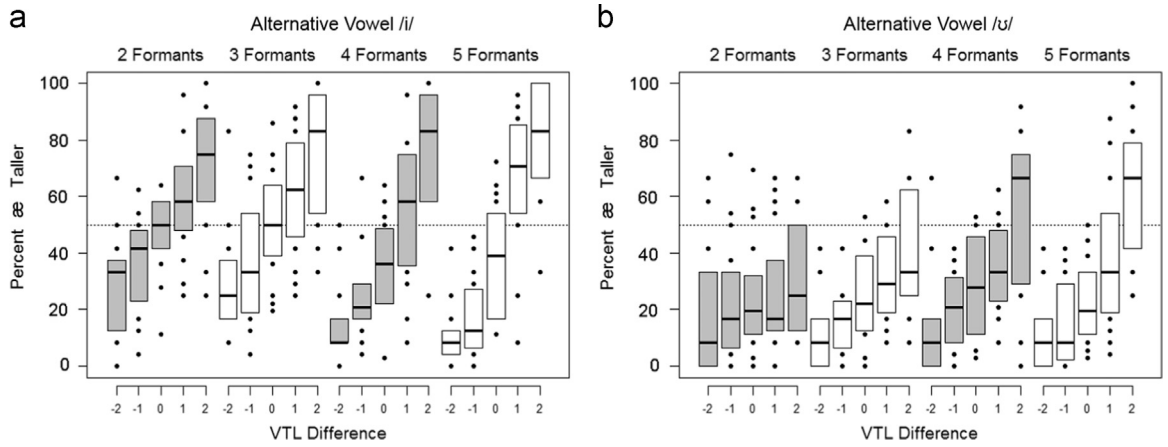

Fig. 6. Points represent the percent trials where /æ/ was selected as taller, averaged within-participant and presented according to the alternative vowel category and formant group. VTL difference refers to the difference in VTL step between the reference vowel and the alternative vowel in the pair. A positive VTL difference that the reference vowel had a longer simulated VTL, which should generally result in the perception of a larger speaker. Boxplots indicate medians and interquartile ranges; whiskers have been omitted so that outliers can be clearly seen.

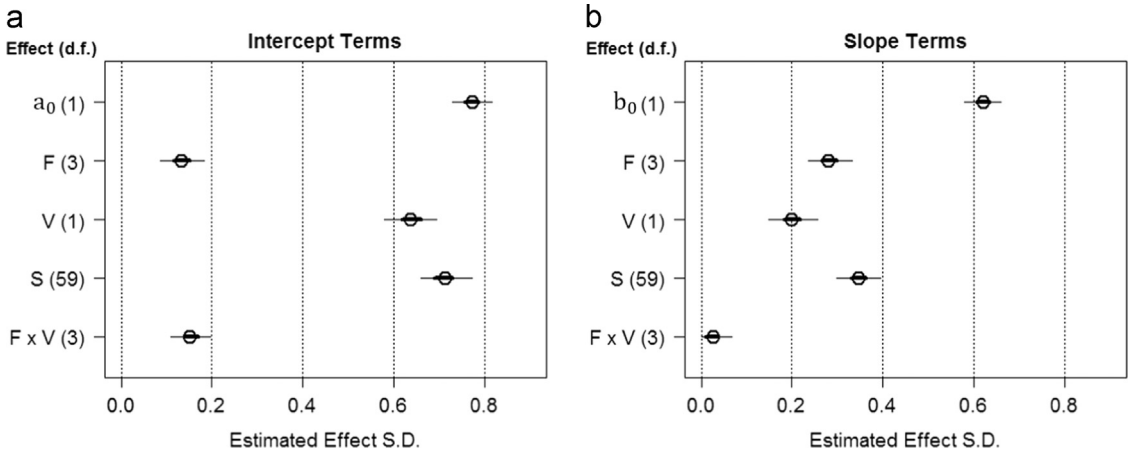

Fig. 7. Results of a Bayesian analysis of variance using the finite-sample standard deviations for the (a) intercept, and (b) slope terms presented in Eqs. (5) and (6). Effects are indicated using the appropriate subscripts. In both panels circles indicate medians, bold lines indicate 50% HDI, and thin lines indicate 95% HDI for posterior distributions.

varies primarily due to the alternative vowel category (α_V) and speaker-specific differences (α_S). Interestingly, whereas in same-phoneme trials individual subject variability had a standard deviation nearly three times that of vowel-category effects (Fig. 5a), in different-phoneme trials these are roughly equal in magnitude. This indicates that there is a good degree of consistency in vowel-category biases relative to the degree of variation between listeners.

Although the effects for number of formants, and the number of formants by vowel category interaction are relatively small, these are both also credibly non-zero (based on the lower extent of their respective HDI). To investigate these effects, the intercepts for different conditions were found by adding together appropriate parameter estimates. For example, the intercept for /i/ for the 4-formant group is equal to: $\alpha_0+\alpha_{F=4}+\alpha_{V=i}+\alpha_{F=4\times V=i}$, while the intercept for /ʊ/ for the 3-formant group is equal to: $\alpha_0+\alpha_{F=3}+\alpha_{V=\upsilon}+\alpha_{F=3\times V=\upsilon}$. The distributions of these reconstructed effects are presented in Fig. 8a.

The purpose of investigating vowel-category biases was first, to see if they exist at all and second, to see if these diminished as more higher-formant information is presented in the vowel. First, it is clear from Fig. 8a that, on average, /æ/ is identified as smaller when compared to either /i/ (mean$=-0.32$; HDI=-0.379, -0.264; 100%<0) or /ʊ/ (mean$=-1.22$; HDI$=-1.29$, -1.16; 100%<0). Furthermore, more higher-formants present in vowels did not result in a reduction of these vowel-category biases. Instead, the higher formants had no effect on the rate at which /ʊ/ was selected as taller compared to /æ/, and more higher-formants resulted in credible *increase* in the bias towards identifying /i/ as taller ($\alpha_{V=i,F=4}-\alpha_{V=i,F=3}$, mean$=0.529$; HDI$=0.367$, 0.686; 100%>0).

Fig. 7b indicates a large non-zero average slope (β_0) that varies primarily according to the vowel category (β_V), the number of formants (β_F), and speaker-specific differences (β_S). Unlike for the intercept terms in 8a, there is no formant group by vowel interaction on slopes indicating that these may be analyzed in terms of independent formant group and vowel category effects. The posterior distributions of these effects, found by adding the appropriate deflection coefficient to the overall slope, β_0, are presented in Fig. 8b and c. Differences in slope arising from the different alternative vowels were investigated by finding the difference between vowel deflection terms ($\beta_{V=i}-\beta_{V=\upsilon}$), which indicated a credibly larger VTL-difference slope for /i/ relative to /ʊ/ (mean$=0.284$; HDI $=0.209$, 0.36; 100%>0), likely arising from the larger overall bias towards identifying /ʊ/ as taller. Credible differences in slopes were found between 2 and 3-formant vowel groups ($\beta_{F=3}-\beta_{F=2}$, mean$=0.14$; HDI$=0.038$ to 0.243; 99.61%>0), as well as between the 3 and 4 formant vowel groups ($\beta_{F=4}-\beta_{F=3}$, mean$=0.178$; HDI$=0.068$, 0.287; 99.94%>0) and the 4 and 5 formant vowel groups ($\beta_{F=5}-\beta_{F=4}$, mean$=0.334$; HDI$=0.214$, 0.461; 100%>0), indicating that listeners will use *F*4 and *F*5 in assessing speaker-size, if available.

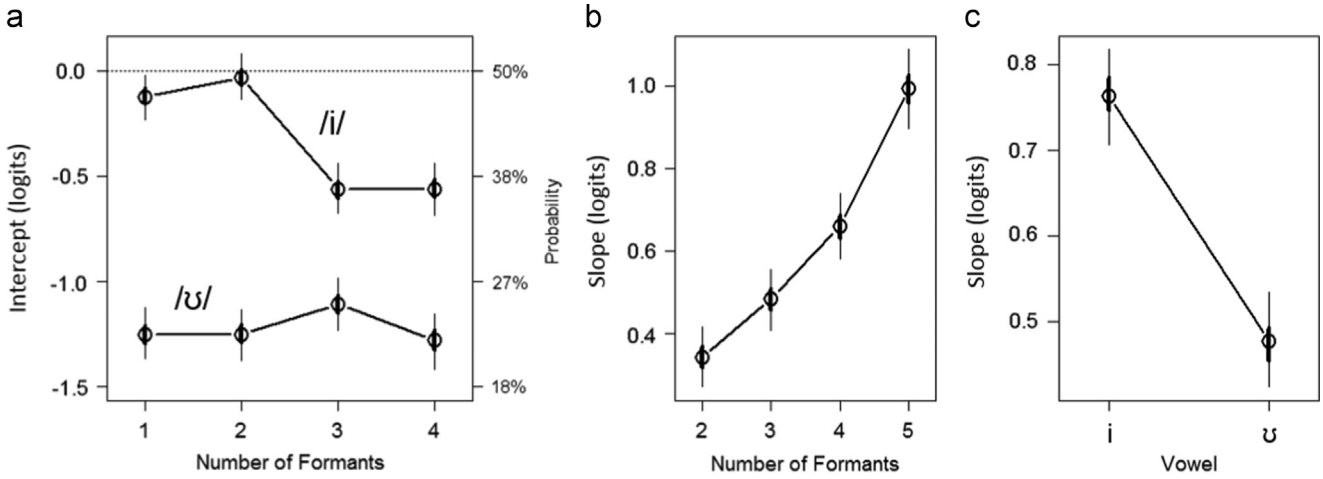

Fig. 8. Posterior distributions of (a) intercept terms for different vowel categories and formant groups, with a different line for each alternative vowel category, (b) average VTL slope for each formant group, and (c) average VTL slopes for each alternative vowel. In all panels, circles indicate means, bold lines indicate 50% HDI, and thin lines indicate 95% HDI for posterior distributions.

4. Discussion

4.1. Phoneme biases in speaker-size perception

The experiment outlined here presented listeners with pairs of stimuli at different simulated vocal-tract length (VTL) levels, and asked listeners to identify the taller speaker in the pair. In this respect, it is very similar in its design to studies previously carried out investigation the perception of speaker size (Rendall et al., 2007; Smith et al., 2005). Crucially, in this experiment linguistic content was not fixed across stimulus pairs, thereby allowing for a direct investigation into the use of spectral information in determining speaker size on a trial by trial basis. If listeners estimate speaker size using a recovered VTL parameter, then size judgments should be based solely on the simulated VTL differences between voices. On the other hand, if listeners use the FFs directly to make these judgments, and not only to inform VTL estimates, then some categories may well be consistently perceived as taller than others.

Results presented in Section 3.2 (Figs. 6 and 8a) indicate listeners showed an overall tendency to identify /i/ as taller than /æ/, although this tendency varied as a function of the number of formants in the vowel (this will be discussed in more detail in Section 4.2.). However, this bias towards /i/ not very large, so that the results in Fig. 6a are generally similar to what would be expected if listeners were using VTL differences between voices to estimate relative size differences. That is, the probability that the reference vowel (/æ/) is selected as taller varies primarily as a function of the VTL difference between the voices, and is strongly influenced by them. Furthermore, the reference vowel is selected as taller in roughly 50% of cases at a VTL difference of 0, and results are roughly symmetrical about the horizontal line at 50%.

In contrast to this, there was a large and persistent tendency to identify /ʊ/ as taller than /æ/, regardless of the VTL differences between the voices. First, consider Fig. 6b which presents a very strong bias towards identifying /ʊ/ as taller than /æ/ when presented at the same VTL level. In these situations these vowels were presented with formant values representing an average male speaker of the dialect, and at the same *f*0 level. If listeners were only using VTL cues to estimate speaker height, there is no reason why there should be such a strong and consistent bias towards selecting /ʊ/ as taller in these cases. Although it is reasonable to expect there might be error in VTL estimation, the biases evident in Figs. 6b and 8a are not indicative of random error, but of persistent and predictable effects that are stable across a large group of sixty listeners. Even stronger evidence of this tendency is evident if one considers results for VTL differences of +2 in Fig. 6b. These correspond to situations where /æ/ had a 16% lower formant scaling than the /ʊ/, suggesting a considerably longer VTL. Even in these cases, /ʊ/ was selected as taller in a majority of cases (53.4% percent).

Rather than respond to VTL cues, listeners appear to use the FFs directly, and simply associate lower formants with larger speakers regardless of the VTL implied by them. Although the results in Fig. 6b may not be explainable solely on the basis of simulated VTL differences between the voices, it seems plausible to consider that listeners might identify /ʊ/ as taller because it has much lower *F*1 and *F*2 frequencies. Similarly, the approximately equal preference for /æ/ and /i/ may be attributable to the fact that although /i/ has a much lower *F*1 than /æ/, this may be offset by its higher *F*2 and *F*3 frequencies. Although this limited stimulus design cannot hope to explain the exact relationship between specific FFs and size perception, the intention of this experiment was simply to establish whether this behavior existed at all.

Taken as a whole, the results presented here offer good evidence that listeners do not solely base speaker-size estimates on the basis of an estimated VTL parameter. Instead, they may be strongly influenced by phoneme-specific formant frequencies in the stimuli presented. The result of this is that listeners may not provide stable size estimates for speakers across vowel categories, and the apparent size difference between two given speakers may vary as a function of the spectral content of the tokens being considered.

4.2. The use of the higher formants

Listeners were divided into groups based on whether they heard 2, 3, 4 or 5 formant vowels. As described in Section 2.2.2, the creation of the stimuli was such that the vowels presented to these groups differed only in the presence/absence of a given number of formants. As a result if, for example, there were credible differences in response patterns between the 3 and 4 formant groups, then this is evidence of

the fact that listeners are using the *F*4 in making speaker-size judgments. On the other hand, if result patterns were roughly equivalent for the two groups, this would indicate that the additional formant was not used by listeners.

Listeners were more likely to associate lower FFs with taller speakers (success rates, Fig. 4b) when more formants were presented, and this improvement continued with the addition of each formant, even up to *F*5 (Fig. 5b). Similarly, the effect of simulated VTL differences in different-phoneme trials became stronger with the addition of both *F*4 and *F*5 (Fig. 8b). In same-phoneme trials, a higher success rate indicates consistent associations between lower FFs (and implied VTLs) and larger speakers, while in different-phoneme trials a larger VTL slope indicates that global differences in the FFs between voices had a stronger effect on responses. From the perspective of using voice FFs in a predictable manner, both of these differences can be considered to reflect improved performance on the part of listeners. As a result, taken together these results offer strong evidence that listeners will use information at least up to *F*5 when making speaker-size judgments, if available.

However, it is not the case that higher formants are necessary in order to make size judgments, as results were highly systematic even for two formant vowels. Furthermore, it is not the case that the presence of more higher-formants reduced phoneme biases. In the case of the /æ/-/ʊ/ comparison, more higher-formants had no effect on the tendency to select /ʊ/. In the case of /æ/-/i/, more higher-formants actually led to the increase of a bias towards selecting /i/ as larger. Although the reason for these preferences are not exactly clear, their existence is not entirely surprising given that listeners appear to make direct use of formant patterns when estimating speaker size (see Section 4.1). In light of this, the addition of formants to a pattern, or the movement of one or more formants in a pattern, may lead to differences in apparent speaker size, regardless of the VTL implied by the formant-pattern as a whole.

Since synthetic vowels were used, and these require higher formants to be explicitly specified, it may be tempting to attribute some of these results to inappropriately specified higher-formant frequencies for the different vowel phonemes used in the experiment. However, as discussed in Section 1.2, there appears to be relatively little between-category variability in the higher formants within-speaker, and what variability there is appears to be overwhelmed by idiosyncratic speaker-characteristics. This suggests that, unlike for *F*1 and *F*2 (and perhaps to a lesser extent *F*3), it is not clear if the higher formants can be 'wrong' for a given phoneme as long as they are plausible values for the speaker. Consequently, if listeners do have strong expectations about the placement of the higher formants for different phonemes (and it is not clear that they do), the higher formants of most real speakers will tend to defy these expectations to a greater or lesser extent. As a result, if deviations from expected frequencies in the higher-formants are a source of variability in speaker-size judgments, then these will occur frequently when size judgments are made for real voices, even within-speaker.

For example, although /i/ and /æ/ appear to have almost identical mean *F*4 values within speaker-class for men, women and children (Hillenbrand et al., 1995, Table V), we might imagine that for whatever reason, listeners in this experiment expected a higher *F*4 for /i/ than for /æ/, for the same speaker. Since in this experiment these vowels were presented with the same *F*4 for a given VTL level, this would have resulted in the *F*4 used for /i/ in this experiment to suggest a slightly larger speaker by being lower than expected, even within VTL level. However, 40% of speakers in Hillenbrand et al. (1995) have lower *F*4 frequencies for /i/ than /æ/, and there is a good amount of variability in the difference between *F*4 across these phonemes, within-speaker in the same dataset (sd = 447 Hz). As a result, if the shifting bias for /i/ for vowels with *F*4 (presented in Fig. 8a) is simply the result of *F*4 being 'too low' then this exact same effect should be present when listeners estimate the size of the nearly half of real human voices that also defy these expectations. Please note that this is not meant to suggest that vowel category biases are driven by mismatches in expectations regarding the higher formants, but simply to note that even if this is the case, then any such effects will also be abundant in the judgments of speaker size from real human voices.

4.3. VTL estimation in size perception

It has been suggested that listeners could use knowledge of the relative formant-patterns for different vowel categories to employ a pattern-correction on observed FFs and estimate speaker VTL from the FFs in this manner (Nearey, 1978; Nearey & Assmann, 2007; Turner, Walters, Monaghan, & Patterson, 2009). Other researchers have made strong claims that the human peripheral auditory system automatically segregates size (VTL) information from speech sounds so that size is actually an independent dimension of sound, and that this information is both available to listeners and the driving force behind speaker-size estimates (Irino & Patterson, 2002; Ives et al. 2005; Turner et al., 2009; Smith et al., 2005, 2007; Smith & Patterson, 2005; Patterson & Irino, 2014). Results indicating substantial vowel-category biases in speaker-size judgments are generally problematic for theories that make strong claims about the availability of a speaker-dependent VTL estimate for human listeners, or regarding the use of these estimates in the determination of speaker size.

The differing levels of successful trials by vowel quality in same-phoneme comparisons (Fig. 5b) is also problematic for accounts where reasonably accurate VTL estimates are the driving force behind speaker size perception. Given that all vowel categories featured identical differences in their FFs for equivalent VTL shifts, it is not clear why there should be consistently differing success rates across categories. To the extent that the assessment of speaker size is based on a VTL estimate, and this is estimated with roughly the same accuracy for different vowel categories, there is no reason to expect an effect for vowel category on the ability to associate lower FFs with a larger speaker. Instead, results suggest that perhaps speaker-size estimation may operate at a low level of perception, without any kind of correction for linguistic patterns or estimation of speaker VTL. In light of this, it may be more accurate to say that listeners are responding to FFs shifts in stimuli even when it might appear that they are responding to VTL differences between speakers, real or synthetic.

The results presented here may go some way towards resolving a long-standing problem in speaker-size perception regarding the strong influence of *f*0 on speaker-size judgments despite the fact that *f*0 is not a good predictor of size in adult speakers. For example, Rendall et al. (2007) state that it is "important to consider why, if *F*0 cues are inherently unreliable markers of adult size, listeners' impressions of size are sometimes biased by them" (1216), and Pisanski et al. (2014) state that "[i]n the absence of a strong physical relationship, the strong perceptual association between *F*0 and size poses a paradox" (95).

If, as it appears, listeners use formants directly when estimating speaker size such that /ʊ/ may sound larger than /æ/ for a single speaker, this would seem to be roughly analogous in its approach to the inappropriate use of *f*0 in determining speaker size for adults. For example, Rendall et al. (2007) report that *f*0 differences as small as 20 Hz can overwhelm relatively large VTL cues in speaker size perception, even though a difference of 20 Hz is a very small relative to the *f*0 range employed by a typical speaker. In both cases, cues that should not be interpreted as strong evidence of differences in size (a 20 Hz difference in *f*0, a linguistically motivated difference in FFs within-speaker) are being used as such, leading to inaccurate size estimates. Taken together, this suggests a general association between

low-frequency acoustic cues and larger perceived speakers, regardless of the more nuanced veridical relationship between acoustic voice parameters and speaker size.

5. Conclusion

The current study investigated the use of spectral information in the perception of speaker size by human listeners. Results indicate that speaker-size estimation is not solely on the basis of a phoneme-independent, speaker-dependent vocal-tract length (VTL) estimate. Instead, it appears that listeners identify speaker size on the basis of the particular FFs contained in the stimuli presented to them. Although this generally will not lead to stable speaker-size estimates, it is well known that these estimates tend to be inaccurate for adult speakers (Bruckert et al., 2006; Collins, 2000; Rendall et al., 2007; Van Dommelen & Moxness, 1995). Therefore, this finding is very much in line with, and may help explain, the general lack of accuracy in size judgments for adult speakers. However, even though the use of spectral cues may be suboptimal such that it will tend to lead to inaccurate speaker-size judgments, there is good evidence that use of the spectral content in vowel sounds is systematic nonetheless. For example, although listeners demonstrated a strong bias towards identifying /ʊ/ as taller than /æ/, regardless of VTL differences, listeners were generally consistent in these biases.

The results presented here suggest that research into speaker-size perception by human listeners would benefit from a shift in focus towards how specific acoustic cues are used by listeners on a trial by trial basis rather than only considering aggregate behavior across a range of stimuli. Collecting size judgments for multiple vowel categories and modeling average judgments is representative of listener behavior if and only if all tokens produced by a single speaker contain the same information (i.e., VTL cues). On the other hand, given that listeners appear to respond to token-specific information, all stimuli cannot be said to contain the same information, and so details regarding listener behavior may be lost when only considering aggregate behavior. Consequently, modeling speaker-size judgments for each given stimulus as a function of the specific characteristics of that stimulus may yield richer information, and a truer representation of how listeners are actually estimating speaker size on any given trial.

Appendix A. Between-speaker variation in formant patterns

Within-phoneme variability between speakers of a single dialect is primarily (but not exclusively) attributable to differences in vocal-tract length (Fant, 1970). The effects of differences in vocal-tract length (VTL) between speakers on the formant-patterns they produce can be modeled as uniform multiplicative increases or decreases (i.e., uniform scaling) of the formant-patterns produced for a given vowel-category, with good accuracy (Barreda & Nearey, 2013; Nearey & Assmann, 2007; Nearey, 1978; Turner et al., 2009). This means that, if two productions of /e/ by two speakers differ by 10% in their *F*1 frequencies, then they are also expected to differ by roughly the same amount in their *F*2 and *F*3 frequencies on average, across all speakers of the dialect. Variation of this kind only requires a single parameter (e.g., VTL) to describe formant-pattern differences between speakers of a dialect since these differences will manifest equally across all formants.

Alternatively, it has been suggested that differences in the oral and pharyngeal cavity lengths between males and females result in non-uniformities in the scaling of formant-patterns between speakers (Fant, 1975). Further, since these non-uniformities are hypothesized to result from the proportionally shorter pharynxes of female speakers, non-uniformities were predicted to vary as a function of vowel height. Under this hypothesis, a different scaling parameter is needed for each formant for each vowel category, for males and females, so that scaling of formant patterns is not uniform across genders and vowel categories. If this were true, it would not be appropriate to discuss differences in formant-patterns using a single parameter (e.g., VTL), nor would it be appropriate to simulate VTL differences by scaling the spectral content of speech sounds up or down by a single parameter. However, no evidence of robustness of this effect, or of its importance to perception exists in the literature. On the contrary, there are good reasons to believe that potential differences in oral/pharyngeal cavity ratios do not deterministically result in non-uniformities in formant patterns between males and females.

For example, Turner et al. (2009) report a reanalysis of the Fitch and Giedd (1999) VTL data and find that the data provides no empirical support for a male–female gender binary in oral/pharyngeal cavity ratios. Instead, they find that this cavity ratio varies continuously as a function of speaker height (and VTL) so that there is a good deal of overlap between males and females (just as there is in height). In fact, this suggests that all speakers will have slightly different oral/pharyngeal cavity ratios so that a large amount of non-uniformities in formant patterns might be expected. Despite this, Turner et al. (2009) report that an investigation of several large formant data sets, including developmental data, reveals that there is no evidence of any non-uniformities in formant patterns. The authors conclude that "the anatomical distinction between the oral and pharyngeal divisions of the vocal tract is immaterial to the acoustic result of speech production. For a given vowel, the tongue constriction is simply positioned where it produces the appropriate ratio of front-cavity length to back-cavity length, independent of the location of the oral-pharyngeal junction" (2379).

We may estimate the inappropriateness of uniform scaling of formant patterns as a working hypothesis by investigating the error incurred in predicting formant patterns when these are constrained to vary uniformly between speaker-classes. Fig. A1 presents variability according to gross speaker type and vowel category for three large data sets, comparing this to the direction of expected variation according to uniform scaling (dotted lines). Although there are many small, random deviations from uniform scaling, there are not systematic deviations that are consistent across all three data sets and, in particular, no systematic deviations from uniform scaling on the basis of vowel height. The lack of a consistent effect across dialects is an important point since the persistence and robustness of non-uniformities in formant patterns across data sets is a central component of the sort of anatomical determinism espoused by the non-uniform scaling hypothesis (i.e., that possible differences in cavity ratios must necessarily manifest in acoustic patterns). In fact, the extent to which uniform scaling seems to hold is striking given that data is taken from 369 different speakers, from very different dialect areas, and several decades apart.

Fig. A2 presents the same data presented in Fig. A1, contrasted with dotted lines that indicate where vowel categories could fall if between-speaker variation within-category were exactly constrained to be according to a uniform scaling factor. These lines pass through

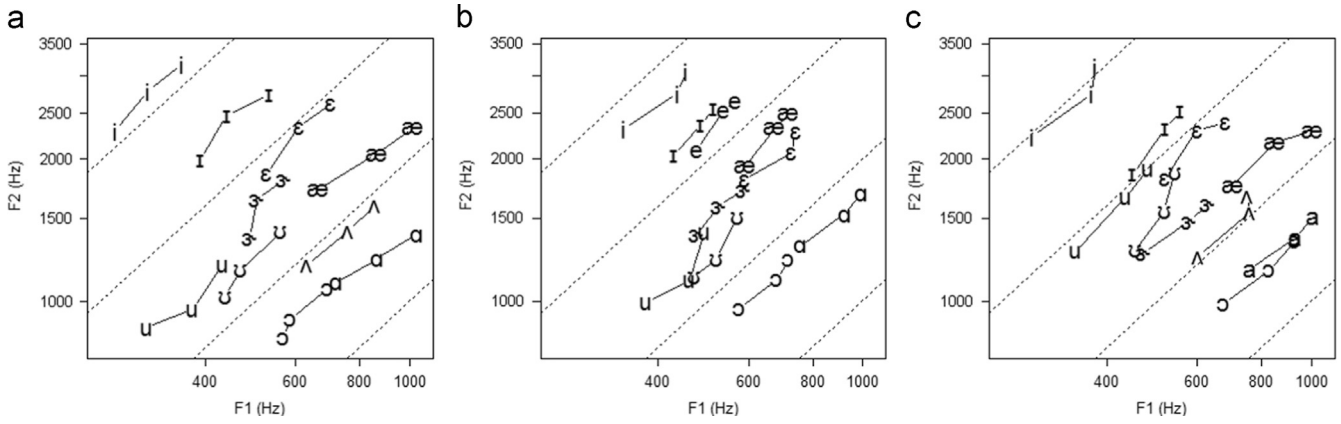

Fig. A1. Average vowel category positions for vowels from three large data sets. (a) Peterson and Barney (1952) contains data from 33 men, 28 women and 15 children of either gender. (b) Syrdal (1985) contains data from 52 men, 51 women and 51 children of either gender. (c) Hillenbrand et al. (1995) contains data from 45 men, 48 women and 46 children of either gender. In each panel, the lowest, left-most instance of each category is the adult male mean, the middle point is the adult female mean, and the top, right-most point is the child mean. Dotted diagonal lines indicate the expected direction of variation according to uniform scaling (i.e., equal multiplicative increases to *F*1 and *F*2). When variation between speaker classes is according to uniform scaling, lines connecting the same vowel category across speaker classes will be parallel to the dotted lines.

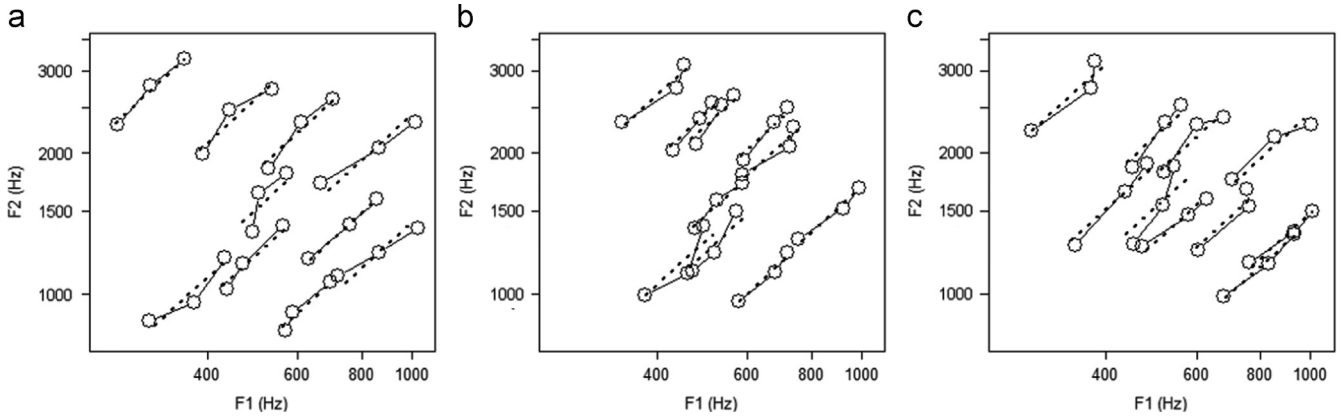

Fig. A2. The same datasets and category means are presented as in Fig. A1, but IPA symbols have been omitted for the sake of legibility. In each panel, the dotted lines pass through the mean *F*1 and *F*2 values for all speakers for each vowel category, but are constrained to have a slope of 1 in a logarithmic space (thereby indicating the direction of uniform scaling of formant patterns). It can be seen that the uniform scaling hypothesis offers a very close approximation of observed formant patterns.

the mean for each vowel category for the three speaker classes, but are constrained to have a slope of one in a log space, meaning they vary along the axis of variation representing equal multiplicative scaling of *F*1 and *F*2.

There is a very close alignment between the observed category mean for each speaker class, and the closest point on the appropriate dotted line. Constraining category means for all speaker classes to vary according to uniform scaling results in an average error in *F*1 and *F*2 of only 2.3%, based on Cartesian distances between observed category means and the nearest point on the dotted lines indicated in Fig. A2. Given that it seems likely that at least some of the deviations from uniform scaling in Fig. A2 are due to noise and/or measurement error and do not represent veridical speaker-class characteristics, it is debatable whether this 2.3% average deviation from uniform scaling should even be considered error. In fact, the magnitude of this error is much smaller than even the amount of variability that can be observed for productions of a single vowel by a single speaker. Peterson and Barney (1952) report two repetitions for each vowel, for each speaker, and these repetitions have an average error of 9.4% in their FFs (based on the Cartesian distance in *F*1 and *F*2 between repetitions).

There is also evidence suggesting the lack of an important perceptual influence for potential non-uniformities in formant-patterns. When creating stimuli for research, it is common practice to simulate VTL differences by linear scaling of the spectral envelope/FFs of stimuli (Assmann et al., 2006; Ives et al., 2005; Rendall et al., 2007; Smith et al., 2005, 2007), for example by using vocoders such as STRAIGHT (Kawahara, Masuda-Katsuse, & de Cheveigné, 1999) or those offered by Praat (Boersma & Weenink, 2001). These more modern processing methods are similar in effect to an old and commonly used recording technique of recording voice actors at one speed, and then playing back the recording at a different speed in order to effect perceived size-changes (Lawson & Persons, 2004). This technique results in uniform scaling of the spectral envelope (and FFs), and has been used in countless films and television shows going back to at least 1939 (Winer, 2012, p. 199). If non-uniform scaling of formant patterns between speakers of different sizes were an important aspect of human speech perception, we should expect to see substantial vowel-quality changes in all of the aforementioned cases, leading to generally less intelligible speech. Instead, the use of the aforementioned techniques appears to have no negative perceptual consequences at all, suggesting that even if non-uniformities in formant patterns exist, they do not play an important role in speech perception. There is also evidence indicating that speech maintains a high degree of naturalness when uniform FF shifts are paired with appropriate *f*0 shifts (Assmann et al., 2006), and that vowel sounds remain highly identifiable when shifted up or down uniformly in FFs (Assmann & Nearey, 2008), even when male vowels are shifted to female ranges and vice versa.

In light of the scant empirical evidence supporting the persistence and robustness of non-uniform scaling in formant patterns between speakers, the lack of any considerations regarding non-uniform scaling in acoustic manipulations of speech (particularly those used in speech research) with no negative consequences for perception, and the lack of any notable error in the prediction of formant patterns according to uniform scaling, it seems reasonable to adopt uniform scaling as a working hypothesis, at least in the absence of more compelling evidence as to the inescapability and importance of non-uniform scaling effects. In fact, the adoption of the uniform scaling hypothesis is in accordance with the main line of research being carried out in the perception of speaker size for the last several decades. Any time a researcher uses linear scaling of the spectral envelope in order to simulate differences in VTL (Assmann et al., 2006; Barreda & Nearey, 2013; Fitch, 1994; Barreda, 2012; Ives et al., 2005; Rendall et al., 2007; Smith et al., 2005, 2007), and any time a researcher uses a single parameter to index speaker spectral characteristics (Collins, 2000; Ives et al., 2005; Pisanski et al., 2014; Rendall et al., 2007; Smith & Patterson, 2005; Smith et al., 2005; Van Dommelen & Moxness, 1995), they are at least tacitly relying on the appropriateness of uniform scaling in formant patterns between speakers of a language, if only as a useful first approximation to variation in formant patterns between speakers of a language.

References

Assmann, Peter F., & Nearey, Terrance M. (2008). Identification of frequency-shifted vowels. *The Journal of the Acoustical Society of America*, *124*(5), 3203–3212. http://dx.doi.org/10.1121/1.2980456.

Assmann, P.F., S. Dembling, T.M. Nearey. (2006). Effects of frequency shifts on perceived naturalness and gender information in speech. In *Ninth international conference on spoken language processing*.

Bachorowski, Jo-Anne, & Owren, Michael J. (1999). Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *The Journal of the Acoustical Society of America*, *106*(2), 1054–1063. http://dx.doi.org/10.1121/1.427115.

Barreda, Santiago (2012). Vowel normalization and the perception of speaker changes: An exploration of the contextual tuning hypothesis. *The Journal of the Acoustical Society of America*, *132*(5), 3453–3464. http://dx.doi.org/10.1121/1.4747011.

Barreda, Santiago, & Nearey, Terrance M. (2013). Training listeners to report the acoustic correlate of formant-frequency scaling using synthetic voices. *The Journal of the Acoustical Society of America*, *133*(2), 1065–1077. http://dx.doi.org/10.1121/1.4773858.

Barreda, S., & Nearey, T. M. (2012). The direct and indirect roles of fundamental frequency in vowel perception. *The Journal of the Acoustical Society of America*, *131*(1), 466–477.

Boersma, Paul, & Weenink, David (2001). *Praat, a system for doing phonetics by computer.*

Bruckert, Laetitia, Liénard, Jean-Sylvain, Lacroix, André, Kreutzer, Michel, & Leboucher, G.érard (2006). Women use voice parameters to assess men's characteristics. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1582), 83–89. http://dx.doi.org/10.1098/rspb.2005.3265.

Collins, Sarah A. (2000). Men's voices and women's choices. *Animal Behaviour*, *60*(6), 773–780. http://dx.doi.org/10.1006/anbe.2000.1523.

Deng, Li, & O'Shaughnessy, Douglas (2003). *Speech processing: A dynamic and optimization-oriented approach*. CRC Press.

Donai, Jeremy J., & Lass, Norman J. (2015). Gender identification from high-pass filtered vowel segments: The use of high-frequency energy. *Attention, Perception, & Psychophysics* (pp. 1–11), 1–11.

Fant, Gunnar (1970). *Acoustic theory of speech production*. Walter de Gruyter.

Fant, Gunnar (1975). Non-uniform vowel normalization. *STL-QPSR*, *16*(2–3), 1–19.

Fitch, William Tecumseh Sherman (1994). *Vocal tract length perception and the evolution of language*. Brown University.

Gelman, Andrew (2005). Analysis of variance—Why it is more important than ever. *The Annals of Statistics*, *33*(1), 1–53. http://dx.doi.org/10.1214/009053604000001048.

Gelman, Andrew (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, *1*(3), 515–534. http://dx.doi.org/10.1214/06-BA117A.

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

Gelman, Andrew, Hill, Jennifer, & Yajima, Masanao (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211.

González, Julio (2004). Formant frequencies and body size of speaker: A weak relationship in adult humans. *Journal of Phonetics*, *32*(2), 277–287. http://dx.doi.org/10.1016/S0095-4470(03)00049-4.

González, Julio (2006). Research in acoustics of human speech sounds: Correlates and perception of speaker body size. *Recent Research Developments in Applied Physics*, *9*, 1–15.

Greisbach, R. (1999). Estimation of speaker height from formant frequencies. *Forensic Linguistics*, *6*(2), 265–277.

Hayakawa S.,& Itakura F. (1995). The influence of noise on the speaker recognition performance using the higher frequency band. In *Proceedings of the international conference on acoustics, speech, and signal processing (ICASSP)* (pp. 321–324). Detroit: IEEE.

Hillenbrand, James, Getty, Laura A., Clark, Michael J., & Wheeler, Kimberlee (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111. http://dx.doi.org/10.1121/1.411872.

Hillenbrand, James M., & Clark, Michael J. (2009). The role of *f*0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics*, *71*(5), 1150–1166. http://dx.doi.org/10.3758/APP.71.5.1150.

Hollien, Harry, Green, Rachel, & Massey, Karen (1994). Longitudinal research on adolescent voice change in males. *The Journal of the Acoustical Society of America*, *96*(5), 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, Toshio, & Patterson, Roy 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*(3), 181–203. http://dx.doi.org/10.1016/S0167-6393(00)00085-6.

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

Joos, M. (1948). Acoustic phonetics. *Language*, *24*(2), 5–136.

Kawahara, Hideki, Masuda-Katsuse, Ikuyo, & de Cheveigné, Alain (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based *F*0 extraction: Possible role of a repetitive structure in sounds1. *Speech Communication*, *27*(3–4), 187–207. http://dx.doi.org/10.1016/S0167-6393(98)00085-5.

Klatt, Dennis H. (1980). Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America*, *67*(3), 971–995.

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

Kruschke, John K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*(7), 293–300. http://dx.doi.org/10.1016/j.tics.2010.05.001.

Kruschke, John K., Aguinis, Herman, & Joo, Harry (2012). The time has come Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, *15*(4), 722–752.

Kruschke, John K., & Vanpaemel, Wolf (2015). Bayesian estimation in hierarchical models. *The Oxford Handbook of Computational and Mathematical Psychology*, *279*.

Labov, William (1972). *Sociolinguistic patterns*, Vol. 4. University of Pennsylvania Press.

Ladefoged, Peter, & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America*, *29*(1), 98–104. http://dx.doi.org/10.1121/1.1908694.

Lass, Norman J., & Brown, William S. (1978). Correlational study of speakers' heights, weights, body surface areas, and speaking fundamental frequencies. *The Journal of the Acoustical Society of America*, *63*(4), 1218–1220. http://dx.doi.org/10.1121/1.381808.

Lawson, Tim, & Persons, Alisa (2004). *The magic behind the voices: A who's who of cartoon voice actors*. Univ. Press of Mississippi.

Miller, James D. (1989). Auditory-perceptual interpretation of the vowel. *The Journal of the Acoustical Society of America*, *85*(5), 2114–2134. http://dx.doi.org/10.1121/1.397862.

Myers, Jerome L., Well, Arnold D., & Lorch, Robert F., Jr. (2010). *Research design and statistical analysis: Third Edition* (3rd ed.). New York, NY: Routledge.

Nearey, Terrance Michael (1978). *Phonetic feature systems for vowels*, Vol. 177. Indiana University Linguistics Club.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, *85*(5), 2088–2113.
Nearey, T. M., & Assmann, P. F. (2007). Probabilistic 'sliding-template'models for indirect vowel normalization. *Experimental Approaches to Phonology*, *246*.
Patterson, Roy D., Irino, Toshio. (2014). Size matters in hearing: How the auditory system normalizes the sounds of speech and music for source size. In A. N. Popper, R. R. Fay (Eds.), *Perspectives on auditory research* (pp. 417–40). Springer.
Peterson, Gordon E. (1961). Parameters of vowel quality. *Journal of Speech and Hearing Research*, *4*(1), 10.
Peterson, Gordon E., & Barney (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175. http://dx.doi.org/10.1121/1.1906875.
Pisanski, Katarzyna, Fraccaro, Paul J., Tigue, Cara C., O'Connor, Jillian J. M., Röder, Susanne, Andrews, Paul W., Fink, Bernhard, DeBruine, Lisa M., Jones, Benedict C., & Feinberg, David R. (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 September.
Plummer, Martyn, et al.. 2003. JAGS: A program for analysis of Bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, 125 pp.). Technische Universit at Wien.
R Core Team (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Rendall, Drew, Vokey, John R., & Nemeth, Christie (2007). Lifting the curtain on the wizard of Oz: Biased voice-based impressions of speaker size. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(5), 1208–1219. http://dx.doi.org/10.1037/0096-1523.33.5.1208.
Rose, Phil (2003). *Forensic speaker identification*. CRC Press.
Rose, Phil, & Clermont, Frantz (2001). A comparison of two acoustic methods for forensic speaker discrimination. *Acoustics Australia*, *29*(1), 31–36.
Smith, David R. R., & Patterson, Roy D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *The Journal of the Acoustical Society of America*, *118*(5), 3177–3186. http://dx.doi.org/10.1121/1.2047107.
Smith, David R. R., Patterson, Roy D., Turner, Richard, Kawahara, Hideki, & Irino, Toshio (2005). The processing and perception of size information in speech sounds. *The Journal of the Acoustical Society of America*, *117*(1), 305. http://dx.doi.org/10.1121/1.1828637.
Smith, David R. R., Walters, Thomas C., & Patterson, Roy D. (2007). Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled. *The Journal of the Acoustical Society of America*, *122*(6), 3628. http://dx.doi.org/10.1121/1.2799507.
Syrdal, A. K. (1985). Aspects of a model of the auditory representation of american english vowels. *Speech Communication*, *4*(1–3), 121–135. ⟨http://doi.org/10.1016/0167-6393(85)90040-8⟩.
Titze, Ingo R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America*, *85*(4), 1699–1707.
Turner, Richard E., Walters, Thomas C., Monaghan, Jessica J. M., & Patterson, Roy D. (2009). A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data. *The Journal of the Acoustical Society of America*, *125*(4), 2374. http://dx.doi.org/10.1121/1.3079772.
Van Dommelen, Wim A., & Moxness, Bente H. (1995). Acoustic parameters in speaker height and weight identification: Sex-specific behaviour. *Language and Speech*, *38*(3), 267–287. http://dx.doi.org/10.1177/002383099503800304.
Vaňková, Jitka (2014). Úspěšnost Rũznỳch Formantovỳch Parametrũ Při Rozlišení Mluvčích. *Acta Universitatis Carolinae Philologica*, *no. 1*, 43–54.
Winer, Ethan (2012). *The audio expert: Everything you need to know about audio*. CRC Press.