



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

Available online 17 December 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 u/), 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 F5) are used by listeners when making speaker-size judgments, though not in a manner consistent with VTL-based speaker-size judgments.

© 2015 Elsevier Ltd. All rights reserved.

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

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 in 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, it is a standard practice to simulate VTL differences between speakers/stimuli by increasing or decreasing stimulus FFs by a single¹ 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.

Download English Version:

<https://daneshyari.com/en/article/1100657>

Download Persian Version:

<https://daneshyari.com/article/1100657>

[Daneshyari.com](https://daneshyari.com)