Journal of Phonetics 63 (2017) 1–18

Contents lists available at ScienceDirect

Journal of Phonetics

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

Research Article

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

Santiago Barreda *

University of California, Davis, United States

ARTICLE INFO

Article history:
Received 24 March 2016
Received in revised form 2 January 2017
Accepted 10 March 2017
Available online 13 April 2017

Keywords:
Vowel perception
Speaker size perception
Speaker characteristics
Speaker normalization

ABSTRACT

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

1. Introduction

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

1.1. Vocal-tract length estimates and size-judgments

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

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

http://dx.doi.org/10.1016/j.wocn.2017.03.002
0095-4470/

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

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

1.2. Vocal-tract estimation in speech perception

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

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

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

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

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