Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled[a]

David R. R. Smith[b]
Centre for Neural Basis of Hearing, Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3EG, United Kingdom and Department of Psychology, University of Hull, Cottingham Road, Hull HU6 7RX, United Kingdom

Thomas C. Walters and Roy D. Patterson
Centre for Neural Basis of Hearing, Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3EG, United Kingdom

(Received 26 September 2006; revised 17 September 2007; accepted 24 September 2007)

A recent study [Smith and Patterson, J. Acoust. Soc. Am. **118**, 3177–3186 (2005)] demonstrated that both the glottal-pulse rate (GPR) and the vocal-tract length (VTL) of vowel sounds have a large effect on the perceived sex and age (or size) of a speaker. The vowels for all of the "different" speakers in that study were synthesized from recordings of the sustained vowels of one, adult male speaker. This paper presents a follow-up study in which a range of vowels were synthesized from recordings of four different speakers—an adult man, an adult woman, a young boy, and a young girl—to determine whether the sex and age of the original speaker would have an effect upon listeners' judgments of whether a vowel was spoken by a man, woman, boy, or girl, after they were equated for GPR and VTL. The sustained vowels of the four speakers were scaled to produce the same combinations of GPR and VTL, which covered the entire range normally encountered in every day life. The results show that listeners readily distinguish children from adults based on their sustained vowels but that they struggle to distinguish the sex of the speaker. © *2007 Acoustical Society of America.* [DOI: 10.1121/1.2799507]

PACS number(s): 43.66.Lj, 43.71.Bp, 43.71.An [RAL]   Pages: 3628–3639

I. INTRODUCTION

Much of the variability in the vowels of men, women, and children arises from characteristic differences in glottal-pulse rate (GPR) (Titze, 1989) and vocal-tract length (VTL) (Fant, 1970; Fitch and Giedd, 1999). GPR is perceived as voice pitch; VTL combines with GPR in the perception of speaker size (Smith and Patterson, 2005). Both GPR and VTL increase with age, and they increase disproportionately for males after puberty. Recent advances in auditory vocoders, such as STRAIGHT (Kawahara and Irino, 2004) and PRAAT (Boersma, 2001), have made it possible to vary the VTL of recorded speech without varying the GPR and vice versa. The ability to vary VTL and GPR while preserving the content of the speech and any other distinctive speaker characteristics has led to a series of studies on the role of GPR and VTL in the perception of vowels (e.g., Assmann and Neary, 2003; Smith *et al.*, 2005), syllables (e.g., Ives *et al.*, 2005), and sentences (e.g., Darwin *et al.*, 2003).

In a recent study, Smith and Patterson (2005) used sustained vowels to determine which of the four responses—man, woman, boy, or girl—would be assigned to vowels with a wide range of combinations of GPR and VTL. The results showed, as expected, that shorter VTLs and higher GPRs lead to the perception that the speaker is a child, and longer VTLs and lower GPRs lead to the perception that the speaker is an adult. The authors drew attention to an apparent anomaly in the data, which was that the voices were heard as women less often than might have been expected, and they pointed out that all of the vowels for all of the "different" speakers in that study (that is, all of the different combinations of GPR and VTL) were synthesized from the speech of a single adult male. The current paper presents a follow-up study to determine how the characteristics of the original speaker's voice affect judgments of the speaker's sex and age, when the GPR and VTL of the voices are controlled. Specifically, the experiment of Smith and Patterson (2005) has been replicated with vowels from four different speakers: The vocoder STRAIGHT was used to synthesize vowels with the same wide range of GPR and VTL values for all four speakers. The vocoder preserves most of the information other than GPR and VTL and enables us to determine whether this extra information affects listeners' ability to discriminate whether the original speaker was a man, woman, boy, or girl. The results show that it is possible to distinguish whether the original speaker was a child or an adult but it is difficult to discern the sex of the original speaker when GPR and VTL are equated. Portions of this work have been presented at several conferences (Smith et al., 2006; Smith et al., 2007a; Smith et al., 2007b).

[a] Portions of this work were presented in "The role of input speaker upon judgements of speaker sex and age," British Society of Audiology, Cambridge, UK, 2006, "Role of glottal-pulse rate, vocal-tract length and original talker upon judgements of speaker sex and age," Acoustical Society of America, Salt Lake City, UT, 2007, and "Judging sex and age: Effect of glottal-pulse rate, vocal-tract length and original talker," International Congress on Acoustics, Madrid, Spain, 2007.

[b] Electronic mail: d.r.smith@hull.ac.uk

The GPR and VTL information in sustained vowels. The length and mass of the vocal folds determine the rate at which the vocal folds open and close. The perceptual marker of GPR is voice pitch; the greater the GPR, the higher the perceived pitch of the voice. There is a strong link between speaker sex and GPR (Darwin, 1871; Morton, 1977). Men have pitches about an octave lower than women primarily because the vocal folds of men are about 60% longer than those of women (Titze, 1989). Voice pitch is a highly salient cue to sex and age because large men have low pitches, young children have high pitches, and adult women lie in the middle (averaging around 105, 260, and 220 Hz, respectively, cf. Huber *et al.*, 1999). Voice pitch can be derived from individual sustained vowels and people are highly sensitive to differences in voice pitch—the just noticeable difference is around 2% (Smith *et al.*, 2005). The sexual dimorphism in GPR is attributable to increased testosterone at puberty in males which stimulates growth in the laryngeal cartilages (Beckford *et al.*, 1985).

Although there are clear intergroup pitch differences *between* men and women, the correlation between GPR and body height *within* a group of adult men, or a group of adult women, is rarely statistically significant, e.g., Lass and Brown (1978), Künzel (1989), Hollien *et al.* (1994), and González (2004). The correlation between GPR and speaker size is also weakened by our use of GPR variation to make prosodic distinctions, such as the rising pitch contour of the interrogative sentence. Indeed, some individuals vary their pitch over an octave during conversation (Hudson and Holbrook, 1982). Thus, in everyday life, GPR provides a strong cue to speaker sex in adults (cf. Bachorowski and Owren, 1999), but it provides only a weak cue to speaker size *within* adult subgroups.

The length and the shape of the vocal tract (VT) causes certain frequencies to be reinforced and others attenuated. The length of the supra-laryngeal VT is highly correlated with speaker height, increasing with age in both sexes (Fitch and Giedd, 1999). The longer the VT, the more the prominent spectral peaks (formants) of speech shift toward lower frequencies (Fant, 1970). Recently, we have shown that small changes in the VTL of vocoded vowels (5%–7%) can be reliably discriminated (Smith *et al.*, 2005; Ives *et al.*, 2005), indicating that speaker size is potentially a perceptually salient aspect of speech. As a child grows between the ages of four and the onset of puberty (around 12), there is a steady increase in VTL with a concomitant decrease in formant frequency. The formant frequencies of adult males decrease by about 30% from their values at age four, while the formant frequencies of adult females decrease by about 20% (Huber *et al.*, 1999). VTL is an important cue to sex and age because it changes with physical size; large adult men have the longest VTLs, children have the shortest VTLs, and adult women have intermediate VTLs (Fitch and Giedd, 1999). The standard deviation for the height of adult men and women is just less than 5%,[1] which is a little less than the just noticeable difference for VTL in adult men. As a result, the correlation between speaker height and formant frequency is predictably weak within the relatively small groups of men or women in published studies (González, 2004; Rendall *et al.*, 2005). Nevertheless, formant-frequency differences clearly distinguish short children from tall adults, and the formant differences can be derived from individual sustained vowels.

In summary, it is clear that GPR and VTL provide potent cues to the main differences in speaker sex and age—that is, whether the speaker is a man, woman, boy, or girl. The question in this study is whether listeners are able to distinguish the sex and age of the original speaker when the study of Smith and Patterson (2005) is rerun with speakers having different sexes and ages.

II. METHOD

Listeners were presented sustained, isolated vowels recorded from four different speakers (adult man and woman, young boy and girl). The vowels were scaled to have the same GPR and VTL values over a large range of GPR and VTL values. Listeners were required to judge whether a boy, girl, man, or woman had spoken each scaled vowel.

A. Stimuli

Examples of the five English vowels (/a:/, /e:/, /i:/, /o:/, /u:/) of an adult man and woman, and a young boy and girl, were recorded using a high-quality microphone (Shure SM58-LCE), with a sampling rate of 48 kHz and a 16-bit amplitude resolution. The vowels were recorded in a sound-attenuating booth to avoid background noise; the microphone was held approximately 5 cm from the chin to maximize the signal to noise ratio. To avoid audible expiration noise, the microphone was held at a point 45° below the horizontal, and the speaker was instructed to pronounce the vowel sounds over, rather than into, the microphone. Speakers were required to utter a series of sustained vowels at a regular relaxed rate (i.e., fifteen /aaaa/ sounds), at a comfortable effort level and at a constant intensity. From these the best five examples were chosen for scaling; that is, vowels were rejected if they had a pitch wobble, jaw articulation noise, lip smacking, or a markedly different pitch from the other examples. For each speaker, five examples of each of the five vowels were selected, giving a total of twenty-five vowel sounds per speaker. Each vowel sound was cut out of the sequence of vowels with care being taken to retain the vowel's natural onset and offset. The age, weight, GPR, height, and estimated VTL (see the following) for each of the four speakers is shown in Table I.

The gain of all the vowel sounds for all speakers was adjusted up or down so that all the vowel sounds had the same rms level prior to scaling. Pilot listening indicated that the vowel sounds had similar loudness.

In order to scale the vowels of the four speakers to the same VTL, it is necessary to estimate the VTL of each speaker. This was done by analyzing the five recorded examples of each of the five vowels /a:/ to /u:/, as spoken by each speaker. The frequencies of formants $F1$ to $F3$ of each vowel were extracted, as a function of utterance time (formant tracks) using PRAAT (Boersma, 2001).[2] The values were found to largely agree with those reported in Hillenbrand *et al.* (1995). These formant tracks were fed into a physical

TABLE I. Physical variables for the four speakers.

Speaker	Age (yr)	Weight (kg)	GPR (Hz)[a]	Height (cm)	VTL (cm)[a]	Height (%)[b]	VTL (%)[b]
Man	24	69.6	108	183	17.6	100	100
Woman	41	68	226	175	14.9	96	85
Girl	9	36	239	143	13.2	78	75
Boy	6	22	256	121	12.5	66	71

[a] Average across all vowels.

[b] Expressed as a percentage normalized to the value for the adult male speaker.

model of formant production tempered by statistical knowledge of VTL and shape variability, and knowledge concerning the error of measurement. The estimates were calibrated against the MRI estimates of vocal-tract length reported by Fitch and Giedd (1999) and the vowel database of Hillenbrand *et al.* (1995). This model performs a factor analysis with a single latent factor of speaker size (Turner *et al.*, 2004). Figure 1 shows estimates of VTL for each speaker, for each of the five vowels, using this model. The scale factors for the speakers were based on the average across all vowels for that speaker. The Appendix describes the model and a calibration test in more detail.

The final step was to create copies of all of the vowels for a wide range of GPR and VTL values for all of the speakers. The scaling of the vowels was performed by STRAIGHT (Kawahara *et al.*, 1999; Kawahara and Irino, 2004). STRAIGHT is a sophisticated vocoder that uses the classical source-filter theory of speech (Dudley, 1939) to segregate GPR information from the spectral-envelope information associated with the shape and length of the vocal tract. Liu and Kewley-Port (2004) have reviewed STRAIGHT and commented favorably on its ability to manipulate formant-related information. STRAIGHT produces a GPR-independent envelope that accurately tracks the motion of the spectral envelope throughout the utterance. Once STRAIGHT has segregated a vowel into a GPR contour and a sequence of spectral-envelope frames, the vowel can be resynthesized with the spectral-envelope dilated or contracted to simulate a change in VTL; the change in VTL for a given scaled vowel is strictly, inversely proportional to the change in spectral-envelope ratio. The GPR dimension (time) can also be expanded or contracted. These operations are largely independent. Utterances recorded from a man can be transformed to sound like a women or a child. The resynthesized utterances are of high quality even when the speech is resynthesized with GPR and VTL values beyond the normal range of speech (provided the GPR is not much greater than the frequency of the first formant, cf. Smith *et al.*, 2005). The duration of all vowels was adjusted to 850 ms within STRAIGHT, by stretching/expanding the signal without altering the pitch or spectral content. The use of STRAIGHT is described in Kawahara and Irino (2004).

The combinations of GPR and VTL used in the experiment are shown in Fig. 2. The values were chosen to encom-

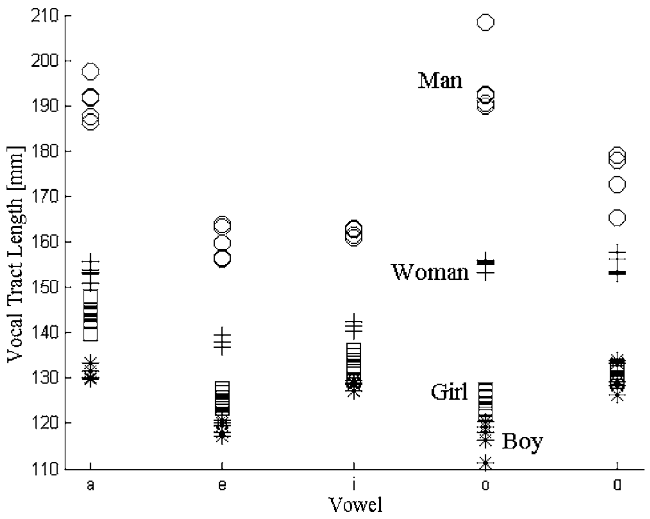

FIG. 1. Estimates of vocal-tract length from formant frequency data using a physical model and a latent variable factor analysis (Turner *et al.*, 2004). At least five examples of each of the vowels /a:/ to /u:/ were analyzed for an adult male (circle), an adult female (plus sign), a young girl (square), and a young boy (asterisk). Details for each of the speakers are provided in Table I.

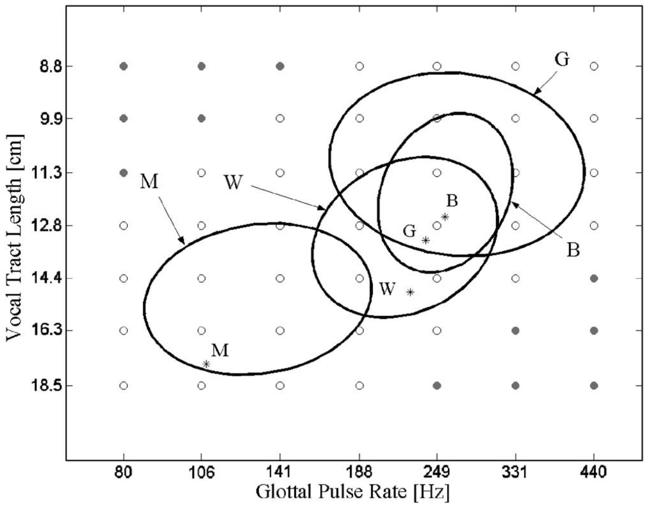

FIG. 2. The open circles show the GPR and VTL combinations of the stimuli used in the sex/age categorization experiment. The GPR values were 80, 106, 141, 188, 249, 331, and 440 Hz. The VTL values were 8.8, 9.9, 11.3, 12.8, 14.4, 16.3, and 18.5 cm. Six conditions in the top-left corner (low GPR and short VTL combinations), and six conditions in the bottom-right corner (high GPR and long VTL combinations), were not presented in the experiment because of their distance from the normal ellipses; these conditions are shown as filled gray circles. The four ellipses show the normal range of GPR and VTL values in speech for men (M), women (W), boys (B), and girls (G), derived from the data of Peterson and Barney (1952). Each ellipse contains 99% of the individuals from the respective category. The asterisks mark the coordinates in the GPR-VTL plane of the four input speakers; man (M), woman (W), boy (B), and girl (G).

pass the range of GPR and VTL encountered in the normal population and to include a large part of the GPR-VTL range employed in Smith and Patterson (2005); GPR varied from 80 to 440 in six, equal, logarithmic steps (seven sample points), and VTL ranged from 18.5 down to 8.8 cm in six, equal, logarithmic steps (seven sample points). The four ellipses show estimates of the normal range of GPR and VTL in speech for men, women, boys, and girls, derived from the Peterson and Barney (1952) vowel database. In each case, the ellipse encompasses 99% of the individuals in the Peterson and Barney data for that category of speaker.[3] Six points in the top-left corner (low GPR and short VTL combinations) and six points in the bottom-right corner (high GPR and long VTL combinations) were not presented because these combinations are unusual and we wished to focus the listeners' attention on normal perception as far as possible.

Listeners were seated in a double-walled, IAC, sound-attenuating booth. The stimuli were played by a 24-bit sound card (Audigy 2, Sound Blaster), through a TDT anti-aliasing filter with a sharp cutoff at 10 kHz and a final attenuator (set at -18 dB), and presented diotically to the listener over AKG K240DF headphones. The sound level of the vowels at the headphones was ~60 dB SPL. The rms level of the vowels was 0.08 (relative to maximum ±1).

B. Procedures

The experiments were performed using a single-interval, four-alternative, forced-choice (4AFC) paradigm. The listener heard scaled versions of five stationary English vowels (/a:/, /e:/, /i:/, /o:/, /u:/), and had to make a judgment about the sex/age of the speaker (man, woman, boy, girl). Sex/age judgments were made by selecting the appropriate button on a response box displayed on a monitor in the booth. The level of the vowel was roved in intensity over a 10 (±5) dB range. Since the judgments are subjective there was no feedback.

A run of judgments consisted of one presentation of each GPR-VTL combination for all five vowels and all four input speakers, presented in a computer-randomized order (a total of 37 GPR-VTL combinations × 5 vowels × 4 input speakers, or 740 trials). For each trial, there were five possible examples of the single vowel that could be played (derived from the five examples of each input vowel for each speaker); the example that was presented was determined pseudorandomly by the computer. Each run took approximately 50 min to complete. Each listener completed five runs in three sessions over a week. Ten listeners participated in the experiments, five male and five female. They ranged in age from 20 to 53 years, and were paid volunteers. All had normal absolute thresholds at 0.5, 1, 2, 4, and 8 kHz.

III. RESULTS

The average results for all listeners are presented in Fig. 3, as four groups of four surface plots. Each group of plots shows the probability of assigning one of the *responses* ("boy," "girl," "man," or "woman") to the sustained vowels of the four *speakers* (boy, girl, man, or woman). For the adult speakers, the distribution of sex and age judgments is very similar across the GPR-VTL plane; that is, the sex of the speaker (man or woman) has relatively little effect on the judgments. Compare the plots for the man and the woman in the bottom row of each judgment group. Similarly, for the children, the distribution of sex and age judgments is largely unaffected by the sex of the speaker (boy or girl); compare the plots for the young boy and girl in the top row of each judgment group. It is also the case that the distribution of "man" responses (bottom-left judgment group) and the distribution of "girl" responses (top-right judgment group) are similar for all four speakers. The effects of original speaker appear mainly in the distribution of "woman" responses (bottom-right judgment group) and the distribution of "boy" responses (top-left judgment group), and they are largely associated with the age (or size) of the speaker. Compare the upper row (boy or girl speaker) with the lower row (man or woman speaker) in the top-left, and bottom-right, judgment groups. With regard to the main experimental question, the results show that when GPR and VTL are controlled, there remains at least one additional cue to the origin of the speaker in sustained vowels, and that cue is more closely associated with the age or size of the speaker (adult *versus* juvenile) than the sex of the speaker (male *versus* female).

Details of the Sex and Age Judgments. The two-dimensional (2D) surface plots in Fig. 3, for the speaker sex and age judgments ("boy," "girl," "man," or "woman"), were constructed as follows: The responses were averaged over the five vowels and all ten listeners, since the pattern of responses was similar for all of the vowels and all of the listeners. Each *group* of four panels shows the probability of the listener assigning the response "boy," "girl," "woman," or "man" (as noted by the group header) to the stimulus vowel, as a function of GPR and VTL. The probability of classification is shown by color, ranging from 0 (dark-blue) meaning "never classified" to 1 (brown-red) meaning "always classified." Within each panel the abscissa is GPR and the ordinate is VTL, both on logarithmic axes. The open circles show the combinations of GPR and VTL presented to the listeners; between these data points, the surfaces have been generated by interpolation. The combinations in the top-left and bottom-right corners of the GPR-VTL plane were not presented because they rarely occur in the population, and as a result, they are omitted from the interpolated surface. The dotted black lines outline regions of the GPR-VTL plane where listeners consistently chose one category of response out of the four available to them. Within these regions, the probability of choosing the given combination of sex and age is greater than 0.5. The four ellipses show estimates of the normal range of GPR and VTL combinations in speech sounds for men, women, boys, and girls (Peterson and Barney, 1952), where each ellipse contains 99% of the individuals from the respective category.

There is one other aspect of the data to note before proceeding to detailed statistical analyses of the results, and that is the listeners' use of the response categories ("man," "woman," "boy," and "girl"), which differs considerably from the distribution of GPR-VTL combinations in the population. In the "man" response group (bottom-left group Fig. 3), the boundary beyond which listeners do not use the

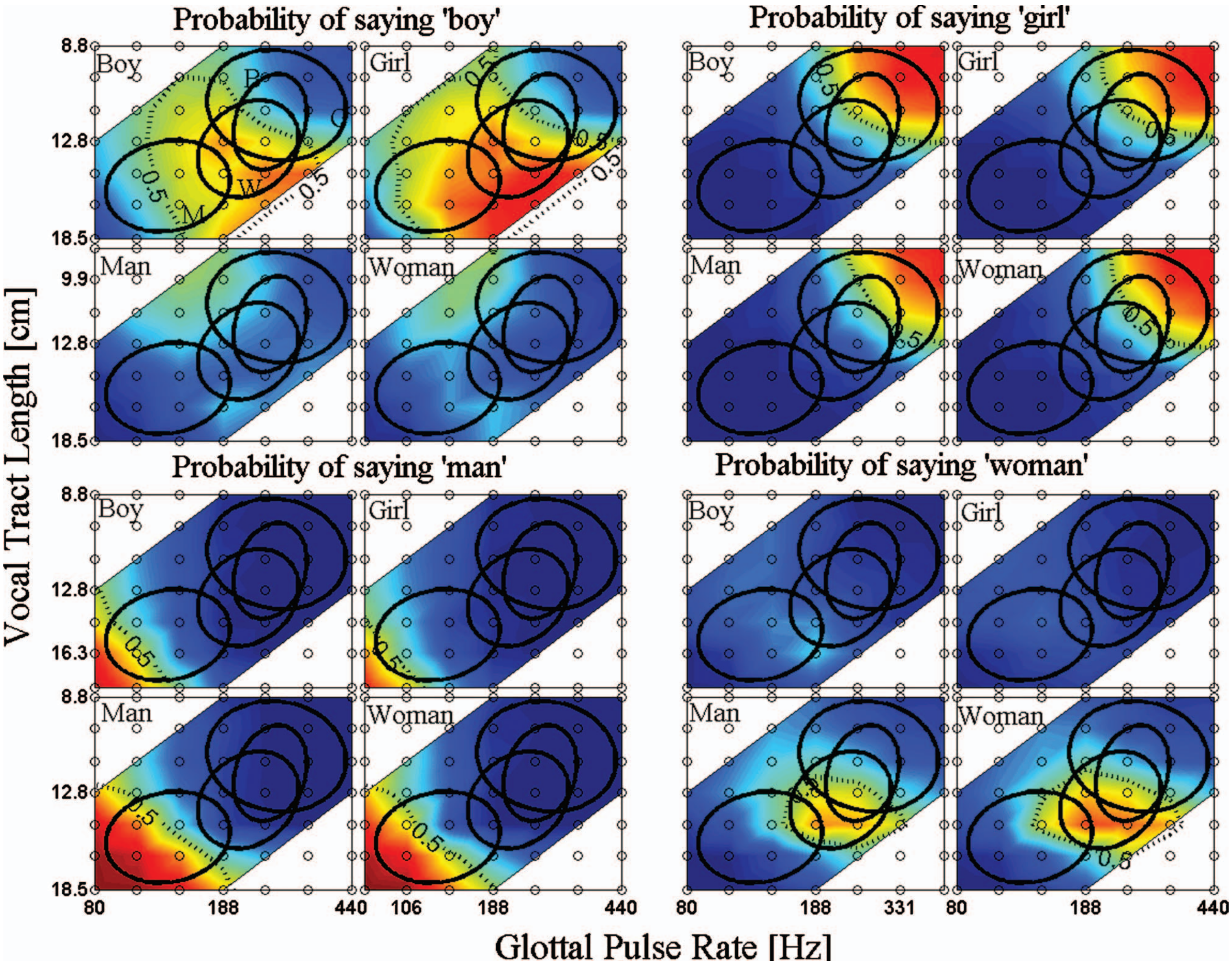

FIG. 3. Sex and age judgments for the four different speakers (young boy, young girl, adult man, and adult woman) averaged over all listeners ($n = 10$). The data are presented as a series of 2D surface plots with color showing probability of assigning a given GPR-VTL combination to one of four perceptual categories (boy, girl, man, or woman). The probability of classification is shown by color, ranging from 0 (dark-blue) meaning "never classified" to 1 (brown-red) meaning "always classified." The data for each perceptual response are shown separately as a group of four panels, where each panel is for a different speaker; top-left quadrant (probability of saying "boy"), top-right quadrant (probability of saying "girl"), bottom-left quadrant (probability of saying "man"), and bottom-right quadrant (probability of saying "woman"). The points in the plane where sex/age judgments were measured are shown by the open circles in each panel; between the data points the surface was derived by interpolation. Within each panel, GPR-VTL combinations in the top-left (low GPR and short VTL) corner and the bottom-right (high GPR and long VTL) corner were not presented. For each GPR-VTL combination, the probabilities from the four panels for a given speaker sum to 1. (Imagine four separate 2D maps stacked vertically and aligned over each other). The data were averaged across all five vowels and all ten listeners, so each sample-point probability is based on 250 trials. The dotted black contour line marks classification threshold, that is, a probability ⩾0.5 of consistently choosing one category out of the four available. The region of GPR-VTL values enclosed by this line defines a region categorized as one particular sex or age. The four ellipses show the range of GPR and VTL in speech for men (M), women (W), boys (B), and girls (G), as derived from the data set of Peterson and Barney (1952).

"man" response is quite abrupt for each original speaker, and the boundary runs across the man ellipse of normal speakers at a point well short of where it might be expected to occur. If listeners were matching to the distribution of speakers in the population, the boundary might be expected to occur near the midline of the intersection of the man and woman ellipses. Similarly, in the "girl" response maps (top-right group Fig. 3), the boundary beyond which listeners do not use the "girl" response is quite abrupt, and it runs across the girl ellipse at a point short of the midline of the intersection of the girl and woman ellipses. So the listeners are not using the distribution of the GPR-VTL combinations in the population to assign their responses. They probably know, at some level, that the distributions for boys and girls overlap, but they assume that the experimenter wants them to be consistent in their use of the responses, and so it does not occur to them that they might distribute their responses probabilistically, in accordance with the overlap in the distributions. As a result, they fail to assign the response "girl" to many combinations of GPR and VTL that might well be produced by girls. It is also the case that vowels with combinations of low GPR and long VTL are all perceived as coming from a large, or very large, person, even when the correct response is "girl," "boy," etc. Similarly, vowels with combinations of high GPR and short VTL are all perceived as coming from a small, or very small, person, independent of the original speaker. So, in these corners of the space, the size aspect of the perception is often at odds with the "correct" response. The listeners were aware that the vowels of the original speakers had all been scaled to all combinations of GPR and VTL.

The situation is different, however, in the central part of the space, where the response is typically "woman" or "boy." Here, the speakers are heard as having sizes within the normal range for humans and there is nothing unusual about them. In this region, the listeners under use the categories "man" and "girl" somewhat, and they overuse the categories "woman" and "boy" somewhat, perhaps because they are, at some level, aware of overusing the "man" and "girl" response for the extreme combinations of GPR and VTL in the corners of the space. In any event, in the central region of the space, listeners use categories that differ in both size and sex ("woman" and "boy"), and there is sufficient ambiguity in the perception of the speaker to allow us to assess the relative effect of the size and sex of the original speaker on the perception. The statistical analyses were designed to quantify the main effects, and to determine whether the statistics confirm that listeners have relatively good information about whether the original speaker was an adult or a child, and at the same time, relatively poor information about the sex of the original speaker.

Three statistical analyses of the data were performed: First, there was an analysis to determine the spatial similarity of the response distributions ("boy," "girl," "man," "woman") between pairs of original speakers (*1. Quantifying the effects of GPR and VTL in sex and age judgments*). Then the details of the speaker effects were explored (*2. Details of speaker effect in sex and age judgments*). Finally, the effects of speaker sex and size were investigated (*3. Role of speaker size and speaker sex in judgments of sex and age*).

1. Quantifying the Effects of GPR and VTL in Sex and Age Judgments. The main effects of GPR and VTL on the distribution of sex and age judgments are shown in Fig. 3. We will begin with the distribution of "man" responses in the bottom-left quadrant; it is similar for all four speakers, inasmuch as vowels with a low GPR and long VTL tend to be categorized as being spoken by a man. The ellipse for adult men shows that this is the natural category to adopt for vowels scaled to these combinations of GPR and VTL. This result replicates Smith and Patterson (2005), who also found that sustained vowels in this region of the GPR-VTL plane are reported as being spoken by men. In Smith and Patterson (2005), the vowels were scaled from a single, adult-male speaker. The present data show that sustained vowels with a low GPR and a long VTL are categorized as being spoken by an adult man, regardless of the source, although the vowels from the boy and girl speakers have to be scaled to more extreme GPR and VTL values than those from the man or woman speakers to produce the same probability of "man" response.

Each of the sixteen panels in Fig. 3 defines a surface of perceptual probability in the GPR-VTL plane. A nonparametric test of spatial association (cf. Ramsden *et al.*, 1999) was used to compare pairs of maps to test whether there is significant overlap. The perceptual maps consist of 37 probability values, each of which represents the judgments of all ten listeners to all five vowels, for a given combination of GPR and VTL. Thus, for each pair of maps compared (map_1 and map_2), there are 37 pairs of probabilities on which to base the comparison. Before comparison, the probabilities were subjected to a hard threshold, such that any point with probability $p \geq 0.5$ was classified as ON; otherwise it was classified as OFF. This results in two maps with binary values which can be compared point for point, using a simple procedure. A four-cell contingency table is constructed which counts corresponding points in each pair of maps which are (i) both ON, (ii) both OFF, (iii) ON in map_1 and OFF in map_2, and (iv) OFF in map_1 and ON in map_2. Each successful match increments the appropriate cell by one. The quantization of the data reduces the power of the test somewhat, but the purpose was just to distinguish the effect of age/size from the effect of sex and, for this purpose, the strong test with conservative criteria is entirely appropriate. "Details of speaker effect in sex and age judgments" to follow deals with the attempt to quantify more subtle effects.

For each pair of maps, the null hypothesis (H_0) that there is *no spatial association* between the two maps is tested. To test if the two maps are spatially associated, we calculate χ^2 from the contingency table using Yates' correction for small cell counts. If the χ^2 value exceeds the critical value for the specified significance level, for one degree of freedom, then we can reject the null hypothesis, and conclude that there is a spatial association between the two maps. The degree of spatial association can be described by the degree of association in the contingency table, known as the contingency coefficient c. It ranges from 0 (meaning the two maps are not correlated at all), to a maximum (meaning the two maps are completely superimposed); the maximum is determined by the number of rows and columns in the contingency table, and for the current measure, the maximum is 0.707 ($1/\sqrt{2}$).

For the response "man," the null hypothesis that there is no association between the perceptual maps generated by different original speakers can be rejected in the majority of cases, with $p < 0.001$ (P(man|b vs P(man|g), P(man|b) vs P(man|m), P(man|b) vs P(man|w), P(man|m) vs P(man|w)). A significant p value indicates the absence of a significant difference between two maps. P(man|b) refers to the conditional probability of responding "man," defined as the matrix of probability responses across the GPR-VTL plane, *given* vowels scaled from a boy speaker. A conservative Bonferroni correction was adopted for the alpha value required for significance, since the four speaker maps were compared in six pair-wise tests; the resulting p value for significance is 0.008 (0.05/6). Using this correction, we *cannot* reject the null hypothesis (that there is *no* association between the response distributions) for two comparisons, the girl speaker *versus* the man speaker [P(man|g) vs P(man|m) is n.s.], and the girl speaker *versus* the woman speaker [P(man|g) vs P(man|w) is n.s.]. The "man" response distribution for vowels scaled from the *girl* speaker is different from the "man" response distributions for vowels scaled from the two adult speakers. Table II shows the results for all of the perceptual-map comparisons. For all speakers except the girl, the distribution of "man" responses seems to be largely determined by the combination of GPR and VTL of the vowel, rather than the sex and age of the original speaker.

The perceptual maps associated with different speakers are also very similar for the "girl" response (top-right quad-

TABLE II. χ^2 test of significance on data collapsed across all five vowels and ten listeners. The first two columns denote the maps being compared; the third column presents the χ^2 value; the fourth column, the significance (or not); and the fifth column, the contingency coefficient. Note that the p value for significance is taken to be 0.008 (=0.05/6); this adopts a conservative Bonferroni correction to compensate for the comparison of four maps in six pair-wise tests for each perceptual response category.

Map_1	Map_2	χ^2	p	c
$P(\text{boy}\mid\text{b})$	$P(\text{boy}\mid\text{g})$	9.7097	n.s. (0.01)	0.46
$P(\text{boy}\mid\text{b})$	$P(\text{boy}\mid\text{m})$	0.0646	n.s.	0.04
$P(\text{boy}\mid\text{b})$	$P(\text{boy}\mid\text{w})$	0.0646	n.s.	0.04
$P(\text{boy}\mid\text{g})$	$P(\text{boy}\mid\text{m})$	0.0191	n.s.	0.02
$P(\text{boy}\mid\text{g})$	$P(\text{boy}\mid\text{w})$	0.0191	n.s.	0.02
$P(\text{boy}\mid\text{m})$	$P(\text{boy}\mid\text{w})$	8.7432	n.s. (0.01)	0.44
$P(\text{girl}\mid\text{b})$	$P(\text{girl}\mid\text{g})$	27.4075	<0.001	0.65
$P(\text{girl}\mid\text{b})$	$P(\text{girl}\mid\text{m})$	27.4075	<0.001	0.65
$P(\text{girl}\mid\text{b})$	$P(\text{girl}\mid\text{w})$	27.4075	<0.001	0.65
$P(\text{girl}\mid\text{g})$	$P(\text{girl}\mid\text{m})$	22.4971	<0.001	0.61
$P(\text{girl}\mid\text{g})$	$P(\text{girl}\mid\text{w})$	22.4971	<0.001	0.61
$P(\text{girl}\mid\text{m})$	$P(\text{girl}\mid\text{w})$	31.7669	<0.001	0.68
$P(\text{man}\mid\text{b})$	$P(\text{man}\mid\text{g})$	16.7733	0.001	0.56
$P(\text{man}\mid\text{b})$	$P(\text{man}\mid\text{m})$	15.1715	0.001	0.54
$P(\text{man}\mid\text{b})$	$P(\text{man}\mid\text{w})$	15.1715	0.001	0.54
$P(\text{man}\mid\text{g})$	$P(\text{man}\mid\text{m})$	8.3159	n.s. (0.01)	0.43
$P(\text{man}\mid\text{g})$	$P(\text{man}\mid\text{w})$	8.3159	n.s. (0.01)	0.43
$P(\text{man}\mid\text{m})$	$P(\text{man}\mid\text{w})$	32.1033	<0.001	0.68
$P(\text{woman}\mid\text{b})$	$P(\text{woman}\mid\text{g})$	[a]	[a]	[a]
$P(\text{woman}\mid\text{b})$	$P(\text{woman}\mid\text{m})$	[a]	[a]	[a]
$P(\text{woman}\mid\text{b})$	$P(\text{woman}\mid\text{w})$	[a]	[a]	[a]
$P(\text{woman}\mid\text{g})$	$P(\text{woman}\mid\text{m})$	[a]	[a]	[a]
$P(\text{woman}\mid\text{g})$	$P(\text{woman}\mid\text{w})$	[a]	[a]	[a]
$P(\text{woman}\mid\text{m})$	$P(\text{woman}\mid\text{w})$	19.0421	0.001	0.58

[a] There are very few "woman" responses for the boy and girl speakers, so in the marked cases the comparison is between two essentially flat planes of no response. It is meaningless to report χ^2 and contingency coefficient values in such cases.

rant group of Fig. 3). Vowels with combinations of high GPRs and short VTLs, which appear in the upper-right corner of each of the GPR-VTL planes, are consistently categorized as being spoken by girls, and there is little effect of original speaker upon the distribution of girl responses. This corner contains the ellipse for girls and the ellipse for boys, but the ellipse for girls extends to higher GPRs and shorter VTLs, so it is arguably the natural category to adopt for the extreme values of GPR and VTL. This result was also reported in Smith and Patterson (2005). We can reject the null hypothesis that there is no spatial association between perceptual maps for the response "girl" ($P(\text{girl}\mid\text{b})$ vs $P(\text{girl}\mid\text{g})$ etc.) with $p<0.001$ (cf. Table II).

However, the null hypothesis for speaker effects cannot be rejected in the case when the response is "boy" or "woman;" the distributions in the upper row are different from those in the lower row for both of these response groups. We will return to this below.

The perceptual maps are very similar for the man and woman speakers (bottom row of each judgment group of Fig. 3). The null hypothesis, that there is no association between corresponding perceptual maps for the man and woman speakers, can be rejected with $p<0.001$ for all perceptual categories except $P(\text{boy}\mid\text{m})$ vs $P(\text{boy}\mid\text{w})$ which is n.s. (0.01) following Bonferroni correction (cf. Table II bottom row of each perceptual group).

The perceptual maps are also very similar for the boy and girl speakers (top row of each perceptual response group of Fig. 3). The null hypothesis, that there is no association between corresponding perceptual maps for the boy and girl, can be rejected with $p<0.001$ for all perceptual categories except $P(\text{boy}\mid\text{b})$ vs $P(\text{boy}\mid\text{g})$ which is n.s. (0.01) following Bonferroni correction (cf. Table II top row of each perceptual group).

In summary, the statistics support the effects observed in Fig. 3; although GPR and VTL have a major effect on the perception of speaker sex and age, they are not the sole determinants of sex and age discriminations based on sustained vowels. Specifically, it is hard to make the sustained vowels of adults sound like those of a boy (top-left quadrant group of Fig. 3), and it is hard to make the sustained vowels of children sound like those of a woman (bottom-right quadrant group of Fig. 3).

2. Details of Speaker Effect in Sex and Age Judgments. The statistical test described above measures *spatial association* between the two perceptual maps. This is a strong test which not only requires that the two maps have similar spread across the GPR-VTL plane, but also requires that the maps are aligned spatially with each other. However, it performs this test globally, which has drawbacks if we wish to investigate more subtle differences between two distributions.

The effect of original speaker upon the judgments can be revealed by comparing the volume of response enclosed by the dotted-line in each panel, across the sixteen panels (Fig. 3), i.e., summing over all sample points with a "probability of response" value $\geqslant 0.5$. Summing only over points with a response $\geqslant 0.5$ means our analysis is focused on parts of the GPR-VTL plane where responses are strong, and where we can be fairly confident about the category to which listeners ascribe the scaled vowel. We do not sum over the *entire* surface within the thresholded region, but only at the sampled points; the response volume is thus discretely rather than continuously defined. This measure gives an accurate and fairer assessment than visual comparisons, because interpolation between sample points in the 2D surface plots sometimes gives a falsely heightened visual impression of the similarities between plots. Table III tabulates these perceptual volumes as a function of the four original speakers. Differences between perceptual responses, as measured by response volume, reveal the effect of speaker upon sex and age judgments. These effects are secondary to those produced by GPR and VTL inasmuch as the simple measure of response volume does not take into consideration the 2D spatial association of the map. Differences between response volumes across subjects can be tested statistically using the t-test for repeated samples.

P(boy): The girl speaker is more likely to elicit the response "boy" than is the boy speaker [$t(9)=-5.028$, $p<0.001$, two-tailed, $P(\text{boy}\mid\text{b})$ vs $P(\text{boy}\mid\text{g})$]. The p value is less than the Bonferroni-corrected significance level of 0.008 (=0.05/6). Obvious differences in "boy" response volume

TABLE III. Response volumes (expressed as percentage of maximum of 37), averaged across all listeners ($n=10$), with standard error of mean (percentage). P(b) means probability of responding "boy" over the sampled GPR-VTL plane, P(g) means probability of responding "girl," etc.

	Boy speaker				Girl speaker				Man speaker				Woman speaker			
	P(b)	P(g)	P(m)	P(w)	P(b)	P(g)	P(m)	P(w)	P(b)	P(g)	P(m)	P(w)	P(b)	P(g)	P(m)	P(w)
x	30.86	21.28	12.99	3.70	41.11	19.78	10.12	2.30	8.41	16.89	26.19	15.66	7.05	20.63	22.00	18.13
s	4.96	2.76	2.59	2.05	4.35	2.78	2.52	1.65	2.66	1.52	1.82	3.36	2.26	1.58	2.08	3.71

between juvenile and adult speakers are subsequently discussed in the paragraph entitled "*P(child) vs P(adult)*." There are no significant differences between adult speakers (P(boy|m) vs P(boy|w).

P(girl): There are no statistical differences in the volumes of the regions that elicit the response "girl" across the different speakers.

P(women): There are no statistical differences in the volume of the regions that elicit the response "woman" across the different speakers. Obvious differences in "woman" response volume between juvenile and adult speakers are subsequently discussed in the paragraph entitled "*P(child) vs P(adult)*."

P(man): The volumes for vowels that elicit the response "man" are greatest for the adult man and woman speakers, and they are substantially smaller for the boy and girl speakers. The difference between adults is significant [$t(9)=4.701$, $p<0.001$, two-tailed, P(man|m) vs P(man|w)], with the adult man speaker being more likely to elicit the response "man" than the woman speaker. The difference between children and adults for P[man] is significant [$t(9)=-5.992$, $p<0.001$, two-tailed, {P(man|b)+P(man|g)} vs {P(man|m)+P(man|w)}], with adult speakers being more likely to elicit an adult response ("man" or "woman") than juvenile speakers.

P(child) vs P(adult): It is possible to scale the vowels of juvenile speakers so that they elicit the response "man" (bottom-left Fig. 3), and to scale the vowels of adult speakers so that they elicit the response "girl" (top-right Fig. 3), *provided* the vowels are scaled to extreme GPR and VTL values. However, for the "boy" and "woman" responses (top-left and bottom-right groups in Fig. 3, respectively), juvenile and adult vowels cannot be equivalently scaled. To quantify this, we compared the response volumes P(child) vs P(adult), where P(child) is defined as P(boy)+P(girl), and P(adult) is defined as P(man)+P(woman), given vowels scaled from juvenile and adult speakers. The null hypothesis is that there is no difference between response volumes P(child) and P(adult) for juvenile and adult speakers. We can reject this for P(child|child) vs P(child|adult) at $p<0.001$ ($t(9)=5.988$, two-tailed), and for P(adult|child) vs P(adult|adult) at $p<0.001$ ($t(9)=-5.561$, two-tailed). Accordingly, it is reasonable to conclude that the response volumes for P(child) and P(adult) differ.

3. Role of Speaker Size and Speaker Sex in Judgments of Sex and Age. The purpose of the following is to explicitly test the hypothesis that speaker *size* has greater power than speaker sex in explaining the distributions of responses. To test the power of speaker size as an explanatory factor, we collapsed the volume-of-response values (sample points $\geqslant 0.5$) across sex (boy and girl speakers versus man and woman speakers), and tested the significance across the four perceptual responses. To test the power of speaker sex as an explanatory factor, we collapsed the volume-of-response values across size (boy and man speakers versus girl and woman speakers), and tested the significance across the four perceptual responses. We tested this using a one-way analysis of variance (ANOVA), where the independent categorical variable was speaker type and the single dependent variable was volume of response.

First, we performed a simple one-way ANOVA with original speaker as four categories of the independent variable, thereby conflating sex and size as factors. The results across the four different perceptual responses are: P(boy) $F(3,36)=20.4$, $p<0.001$, $\eta^2=0.63$; P(girl) $F(3,36)=0.74$, $p=\text{n.s.}$, $\eta^2=0.06$; P(man) $F(3,36)=10.99$, $p<0.001$, $\eta^2=0.48$; P(woman) $F(3,36)=8.23$, $p<0.001$, $\eta^2=0.41$. The correlation ratio, η^2, is a measure of the size of the effect. It is calculated as a proportion of the between sum of squares (i.e., that part of the variation in the data attributable to the independent variable), to the total sum of squares (i.e., that part of the variation in the data attributable to the independent variable plus all other factors), giving $\eta^2=(SS_{\text{bet}}/SS_{\text{tot}})$. Having original speaker as four categories of the independent variable, thereby conflating sex and size as factors, accounts for between ~40% and 65% of the variance in the data when the data are expressed as response volumes.

For speaker size as a factor, we performed a one-way ANOVA with speaker collapsed over sex, i.e., boy and girl speaker versus man and woman speaker. The results across the four different perceptual responses are: P(boy) $F(1,18)=30.66$, $p<0.001$, $\eta^2=0.63$; P(girl) $F(1,18)=0.34$, $p=\text{n.s.}$, $\eta^2=0.02$; P(man) $F(1,18)=15.99$, $p<0.001$, $\eta^2=0.47$; P(woman) $F(1,18)=12.59$, $p=0.002$, $\eta^2=0.41$. Size as an independent variable (disregarding sex) still accounts for between ~40% and 65% of the variance in the data expressed as response volumes.

For speaker sex as a factor, we performed a one-way ANOVA with speaker collapsed over size, i.e., boy and man speaker versus girl and woman speaker. The results across the four perceptual responses are: P(boy) $F(1,18)=1.16$, $p=\text{n.s.}$, $\eta^2=0.06$; P(girl) $F(1,18)=0.25$, $p=\text{n.s.}$, $\eta^2=0.01$; P(man) $F(1,18)=1.61$, $p=\text{n.s.}$, $\eta^2=0.08$; P(woman) $F(1,18)=0.03$, $p=\text{n.s.}$, $\eta^2=0.00$. Sex as an independent variable (disregarding size) does not have the power necessary to explain any differences in the data expressed as response volumes.

In summary, speaker size has greater power than speaker sex in accounting for the variance in our data. This can be appreciated informally by inspecting the figure; if one merges, by eye, the upper and lower panels by *column* for each response group in Fig. 3, then the remaining two panels in each response group would be very similar. Having collapsed over size, comparison between sexes shows little difference. Alternatively, if one merges, by eye, the upper and lower panels in each response group by *row*, the remaining two panels in each response group are very different. Having collapsed over sex, the comparison between sizes reveals substantial differences.

IV. DISCUSSION

Previous research intended to identify the acoustic properties of male and female voices responsible for the perception of gender has used a variety of methods such as statistical clustering and perceptual categorization (e.g., Childers and Wu, 1991; Coleman, 1976; Whiteside, 1998; Wu and Childers, 1991; Bachorowski and Owren, 1999; Schwartz and Rine, 1968; Ingemann, 1968; Lass *et al.*, 1976). The general conclusion from these studies is that the main acoustic variables affecting perception of gender are GPR and VTL, although the relative importance of the two factors is moot in previous research. Motivated by recent work on the perception of auditory size (Smith *et al.*, 2005; Ives *et al.*, 2005; Turner *et al.*, 2006), Smith and Patterson (2005) measured the interaction of GPR and VTL in judgments of age/size, as compared with judgements of sex. In that study, however, all of the "different" (different GPR and VTL combinations) speakers were created from vowels recorded from a single, adult male speaker. The current study was intended to determine how the pattern of response observed in Smith and Patterson (2005) would vary if the original speaker were an adult woman, a young boy, or a young girl, as opposed to an adult man,

The distributions of sex and age judgments across the GPR-VTL plane are very similar for the man and woman speakers (bottom row of each judgment group of Fig. 3), and separately, very similar for the boy and girl speakers (top row of each judgment group of Fig. 3). The "man" distributions (bottom-left quadrant of Fig. 3) and the "girl" distributions (top-right quadrant of Fig. 3) are largely independent of the sex and age of the original speaker. There are, however, clear differences in the use of the "boy" response (top-left quadrant group of Fig. 3) and the use of the "women" response (bottom-right quadrant group of Fig. 3). The results suggest that speaker *size* is more important than speaker *sex* in determining whether vowels are judged to come from a boy, girl, man, or woman (cf. the last section of Sec. III).

In a recent paper, Assmann *et al.* (2006) presented perceptual judgments of voice gender where the speech of an adult male and an adult female speaker had been manipulated to have the same fundamental and formant frequencies. They found that adult female voices can be made to sound like adult male voices and vice versa. However, they did not use children's voices so there are no data concerning the child/adult distinction.

After controlling for GPR and VTL, there remain several interesting effects of original speaker. First, the girl speaker is more likely to be assigned the response "boy" than is the boy speaker, especially for the more extreme values of GPR and VTL. In point of fact, the girl was a little taller and older than the boy, and her GPR and VTL values (239 Hz, 13.2 cm) indicate a slightly larger person than those of the boy (256 Hz, 12.5 cm). The results of the current experiment suggest that, at least for sustained vowels, it is largely size that determines how the voice of a juvenile will be heard. In the current experiment, when listeners can hear that the voice comes from a juvenile speaker, and there is no contextual information to indicate the sex of the speaker, they assign the response "boy" to voices with longer vocal tracts and lower pitches and "girl" to shorter vocal tracts and higher pitches. It might well be that this is a general bias in the population responses. Moreover, it may be the case that, in the absence of contextual information, there is a bias in the perception of speaker sex, with listeners actually hearing the voices of larger children as boys and the voices of smaller children as girls. However, it would take further experiments to confirm such a hypothesis.

Second, the results show that it is hard to make the sustained vowels of children sound enough like a woman to elicit the response "woman," yet if the vowels of children are scaled to extreme values of GPR and VTL, as occurs in the bottom-left corner of the GPR-VTL plane, they are assigned the response "man." Informal listening indicates that this is a true perceptual effect and that the extreme GPR and VTL values override more subtle cues in this region, causing us to conclude not only that the voice comes from a very large person, but also that the person is male. Again, however, it would take further experiments to confirm such a hypothesis. Note, however, that the vowels of the young boy and girl have to be driven to more extreme GPR and VTL values than those of the adult speakers before listeners assign them the response "man."

Beyond GPR and VTL. Scaling all of the voices from one original speaker, as in Smith and Patterson (2005), meant that certain characteristics of the vowels were fixed regardless of VTL. This is probably not typical of the population of human voices as a whole (Fant, 1966, 1975; Diehl *et al.*, 1996). There is an important anatomical difference between children and adults; children have proportionately larger heads relative to their body size than do adults. As a result, the ratio of oral cavity length to pharyngeal cavity length is greater in children than it is in adults (Fant, 1966). Figure 4 shows how oral and pharyngeal length grow as a proportion of VTL as children mature into adults (redrawn from Turner *et al.*, 2004). Since oral/pharyngeal length ratio (OPR) changes markedly with age, it is reasonable to hypothesize that the changes in OPR might produce changes in formant ratios that could account for the effects of size observed in our data; that is, the formant ratios for a given vowel might be somewhat different in children and adults.

Accordingly, we calculated the $F2/F1$ and $F3/F1$ formant ratios for the sustained vowels of our four speakers; the average ratios are plotted, as a function of speaker height, in Fig. 5. The $F2/F1$ ratio increases marginally as speaker

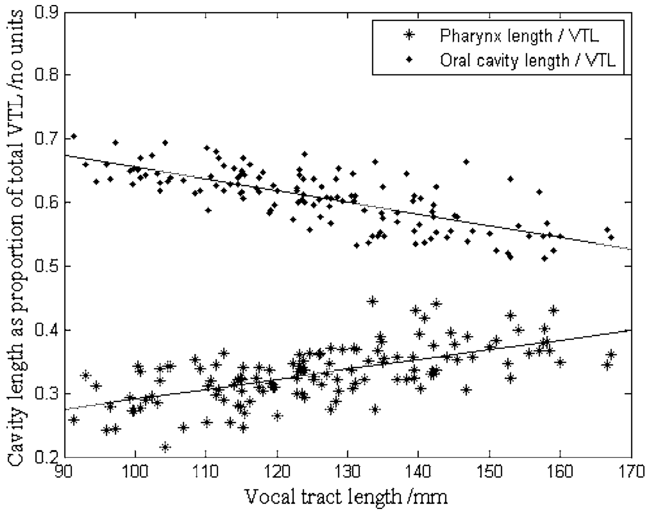

FIG. 4. Oral and pharyngeal cavity lengths, expressed as a proportion of total VTL, and plotted as a function of speaker VTL (redrawn from Turner *et al.*, 2004). The cavity lengths were derived from MRI measurements made by Fitch and Giedd (1999). The lines show the best-fitting linear regression.

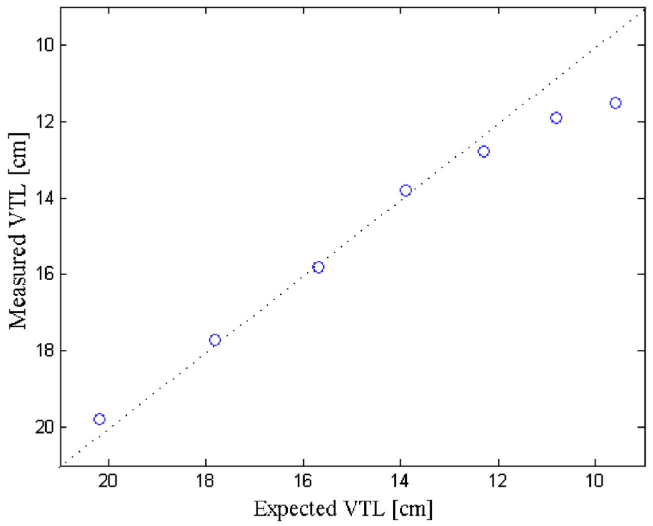

FIG. 6. (Color online) Results of a calibration test involving the scaling applied by STRAIGHT and the physical model (cf. Turner *et al.*, 2004) used to derive VTL from a vowel sound.

height increases while the $F3/F1$ ratio increases somewhat more. The error bars show the standard deviations across vowel; the standard deviations for the individual vowels are much smaller—on the order of the size of the symbols for the means. Since it is not obvious which measure of variability determines discrimination performance, and since there are only four individuals in this study, the data have to be interpreted with caution, but if the results are representative, it is clear that the change in the $F2/F1$ ratio is far less than would be predicted by the nonuniform growth of the OPR, and that it would be difficult to use the information given the variability. It seems more likely that, in attempting to preserve vowel identity, humans vary the position of the tongue constriction to counteract the potential effects of nonlinear growth, and so maintain the $F2/F1$ ratio for individual vowels. It is also possible that there is a limit to the adaptation process, and that it is not possible to preserve both the $F2/F1$ ratio and the $F3/F1$ ratio, which would account for the larger changes in $F3/F1$ with age. In any event, it seems that it is more likely that it is the $F3/F1$ ratio that is the basis of the age effect observed in our data, and if this is the result of the nonuniform growth of OPR the connection is complex.

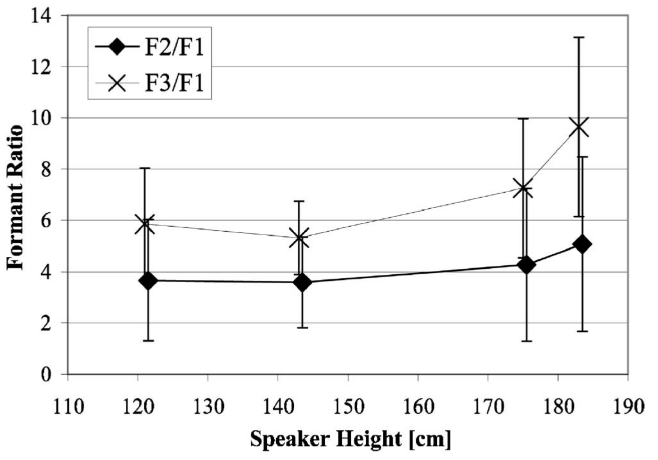

FIG. 5. $F2/F1$ and $F3/F1$ formant ratios of the vowels of the four speakers in the study. The speakers are boy, girl, woman, and man in ascending height along the abscissa. The formant ratios are the average of all five examples of each of the five vowels for each speaker. The formant frequencies were calculated from the middle 10% of each formant track, as extracted by PRAAT (Boersma, 2001). The error bars show ±1 s.d. across vowels for each speaker.

Finally, it is important to remember that the research discussed in this paper deals with acoustic information about speaker sex and age in *sustained vowels*, rather than in sentences or running speech. It may well be that other cues in natural speech, such as increased articulation, would lead to a somewhat different pattern of results.

V. SUMMARY AND CONCLUSIONS

Listeners were presented isolated, sustained vowels recorded from four different speakers (young boy and girl, adult man and woman). The vowels were scaled to produce the same range of GPR and VTL values in each case. Listeners were required to discriminate whether the original speaker was a boy, girl, man, or woman. The results show that, for adult speakers, the distribution of responses across the GPR-VTL plane is largely independent of the sex and the age of the original speaker (top row in each group of Fig. 3). The results also show that, for juvenile speakers, the distribution of responses across the GPR-VTL plane is largely independent of the sex and age of the original speaker (bottom row in each group of Fig. 3). Where differences exist, as in the distribution of "boy" responses (top-left response group of Fig. 3), and "woman" responses (bottom-right response group of Fig. 3), they arise when scaling from adult speakers to juvenile speakers and vice versa. The results show that listeners readily distinguish whether the original speaker was a child or an adult, based on their sustained vowels, but they find it difficult to distinguish the sex of the original speaker.

ACKNOWLEDGMENTS

This research was supported by the UK Medical Research Council (G9900369 and G0500221), the German Volkswagen Foundation (VWF 1/79 783), and the European Office of Aerospace Research and Development (FA8655-05-1-3043).

APPENDIX

Estimates of vocal-tract length were derived from vowel sounds using a physical model and latent variable factor analysis (Turner *et al.*, 2004; cf. Fig. 1 this paper). These estimates were subsequently used to scale different-sized speakers with STRAIGHT (Kawahara *et al.*, 1999; Kawahara and Irino, 2004) to produce matching GPR and VTL values for the four speakers. We report an additional calibration test that feeds the scaled values after manipulation in STRAIGHT back into our physical model to gain an estimate of VTL for the new scaled vowels (Fig. 6). VTL values are plotted as expected VTL (abscissa) versus measured VTL (ordinate). Expected VTL is the value we expect after manipulation in STRAIGHT. Measured VTL is the value measured by our physical model (Turner *et al.*, 2004). For the most part, values fall along the positive diagonal, indicating that VTL values have been correctly scaled. Errors arise for very short VTLs; these are attributable to errors in the ability to accurately extract formant frequencies from these sounds.

[1]Health Survey for England 2004, representative sample of 2436 adult men and 3311 adult women. Average height adult men 1750 mm with a standard deviation of 89.93 mm, and average height adult women 1612 mm with a standard deviation of 69.05 mm [http://www.ic.nhs.uk/pubs/hlthsvyeng2004upd] (last viewed 17 September 2007).

[2]Boersma, P., and Weenik, D. (**2005**). "Praat: doing phonetics by computer" (Version 4.4.30) [Computer program from http://www.praat.org/] (last viewed 17 September 2007).

[3]The set of formant values for each of the 76 speakers in the classic study of Peterson and Barney (1952) were converted to VTL values using the VTL data that Fitch and Giedd (1999) extracted from magnetic resonance images of a large population of subjects. Each ellipse represents the mean±3 s.d. for each category of speaker. The calibration details are presented in Turner *et al.* (2004); this poster can be found at http://www.pdn.cam.ac.uk/groups/cnbh/research/posters_talks/BSA2004/TWPBSA04.pdf. (last viewed 17 September 2007).

Assmann, P. F., Dembling, S., and Nearey, T. M. (**2006**). "Effects of frequency shifts on perceived naturalness and gender information in speech," in *Proceedings of the Ninth International Conference on Spoken Language Processing*, Pittsburgh, PA.

Assmann, P. F., and Neary, T. M. (**2003**). "Frequency shifts and vowel identification," in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain.

Bachorowski, J., and Owren, M. J. (**1999**). "Acoustic correlates of talker sex and individual talker sex identity are present in a short vowel segment produced in running speech," J. Acoust. Soc. Am. **106**, 1054–1063.

Beckford, N. S., Rood, S. R., and Schaid, D. (**1985**). "Androgen stimulation and laryngeal development," Ann. Otol. Rhinol. Laryngol. **94**, 634–640.

Boersma, P. (**2001**). "Praat, a system for doing phonetics by computer," Glot International **5**, 9/10, 341–345.

Childers, D. G., and Wu, K. (**1991**). "Gender recognition from speech. II Fine analysis," J. Acoust. Soc. Am. **90**, 1841–1856.

Coleman, R. O. (**1976**). "A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice," J. Speech Hear. Res. **19**, 168–180.

Darwin, C. (**1871**). *The Descent of Man and Selection in Relation to Sex* (Murray, London).

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**, 2913–2922.

Diehl, R. L., Lindbolm, B., Hoemeke, K. A., and Fahey, R. P. (**1996**). "On explaining certain male-female differences in the phonetic realization of vowel categories," J. Phonetics **24**, 187–208.

Dudley, H. (**1939**). "Remaking speech," J. Acoust. Soc. Am. **11**, 169–177.

Fant, G. (**1966**). "A note on vocal tract size factors and non-uniform F-pattern scalings," STL-QPSR **4**, 22–30.

Fant, G. (**1970**). *Acoustic Theory of Speech Production*, 2nd ed. (Mouton, Paris).

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

Fitch, W. T., and Giedd, J. (**1999**). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," J. Acoust. Soc. Am. **106**, 1511–1522.

González, J. (**2004**). "Formant frequencies and body size of speaker: A weak relationship in adult humans," J. Phonetics **32**, 277–287.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.

Hollien, H., Green, R., and Massey, K. (**1994**). "Longitudinal research on adolescent voice change in males," J. Acoust. Soc. Am. **96**, 3099–3111.

Huber, J. E., Stathopoulos, E. T., Curione, G. M., Ash, T., and Johnson, K. (**1999**). "Formants of children, women and men: The effects of vocal intensity variation," J. Acoust. Soc. Am. **106**, 1532–1542.

Hudson, A., and Holbrook, A. (**1982**). "Fundamental frequency characteristics of young black adults: Spontaneous speaking and oral reading," J. Speech Hear. Res. **25**, 25–28.

Ingemann, F. (**1968**). "Identification of the speaker's sex from voiceless fricatives," J. Acoust. Soc. Am. **44**, 1142–1144.

Ives, D. T., Smith, D. R. R., and Patterson, R. D. (**2005**). "Discrimination of speaker size from syllable phrases," J. Acoust. Soc. Am. **118**, 3816–3822.

Kawahara, H., and Irino, T. (**2004**). "Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Boston, MA), pp. 167–180.

Kawahara, H., Masuda-Kasuse, I., and de Cheveigne, A. (**1999**). "Restructuring speech representations using pitch-adaptive time-frequency smoothing and instantaneous-frequency-based F0 extraction: Possible role of repetitive structure in sounds," Speech Commun. **27**, 187–207.

Künzel, H. J. (**1989**). "How well does average fundamental frequency correlate with speaker height and weight?," Phonetica **46**, 117–125.

Lass, N. J., and Brown, W. S. (**1978**). "Correlational study of speakers' heights, weights, body surface areas and speaking fundamental frequencies," J. Acoust. Soc. Am. **63**, 1218–1220.

Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., and Bourne, V. T. (**1976**). "Speaker sex identification from voiced, whispered, and filtered isolated vowels," J. Acoust. Soc. Am. **59**, 675–678.

Liu, C., and Kewley-Port, D. (**2004**). "STRAIGHT: A new speech synthesizer for vowel formant discrimination," ARLO **5**, 31–36.

Morton, E. S. (**1977**). "On the occurrence and significance of motivation-structural rules in some bird and mammal sounds," Am. Nat. **111**, 855–869.

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

Ramsden, B. M., Hung, C. P., and Roe, A. W. (**1999**). "Activation of illusory contour domains in macaque area V2 is accompanied by relative suppression of real contour domains in area V1," Abstr. Soc. Neurosci. **25**, 2060.

Rendall, D., Vokey, J. R., Nemeth, C., and Ney, C. (**2005**). "Reliable but weak voice-formant cues to body size in men but not women," J. Acoust. Soc. Am. **117**, 2372.

Schwartz, M. F., and Rine, H. E. (**1968**). "Identification of speaker sex from isolated, whispered vowels," J. Acoust. Soc. Am. **44**, 1736–1737.

Smith, D. R. R., and Patterson, R. D. (**2005**). "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age," J. Acoust. Soc. Am. **118**, 3177–3186.

Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H., and Irino, T. (**2005**). "The processing and perception of size information in speech sounds," J. Acoust. Soc. Am. **117**, 305–318.

Smith, D. R. R., Walters, T. C., and Patterson, R. D. (**2007a**). "Role of glottal-pulse rate, vocal-tract length and original talker upon judgements of speaker sex and age," J. Acoust. Soc. Am. **121**, 3135–3136.

Smith, D. R. R., Walters, T. C., and Patterson, R. D. (**2007b**). "Judging sex and age: Effect of glottal-pulse rate, vocal-tract length and original speaker," in *Proceedings of the 19th International Congress on Acoustics ICA2007*, Madrid, Spain. Special issue Revista de Acústica **38**, PPA-090-010.

Smith, D. R. R., Walters, T. C., Walland, K., and Patterson, R. D. (**2006**). "The role of input speaker upon judgements of speaker sex and age," paper presented at British Society of Audiology, Cambridge, p. 43.

Titze, I. R. (**1989**). "Physiologic and acoustic differences between male and female voices," J. Acoust. Soc. Am. **85**, 1699–1707.

Turner, R. E., Al-Hames, M. A., Smith, D. R. R., Kawahara, H., Irino, T., and Patterson, R. D. (**2006**). "Vowel normalisation: Time-domain processing of the internal dynamics of speech," in *Dynamics of Speech Production and Perception*, edited by P. Divenyi, S. Greenberg, and G. Meyer (IOS Press, Amsterdam), pp. 153–170.

Turner, R. E., Walters, T. C., and Patterson, R. D. (**2004**). "Estimating vocal tract length from formant frequency data using a physical model and a latent variable factor analysis," paper presented at British Society of Audiology, UCL, London, p. 61, http://www.pdn.cam.ac.uk/groups/cnbh/research/posters_talks/BSA2004/TWPBSA04.pdf (Last accessed 9/17/07).

Whiteside, S. P. (**1998**). "Identification of a speaker's sex from synthesized vowels," Percept. Mot. Skills **86**, 595–600.

Wu, K., and Childers, D. G. (**1991**). "Gender recognition from speech. P I. Coarse analysis," J. Acoust. Soc. Am. **90**, 1828–1840.