Identification of speaker sex from one vowel across a range of fundamental frequencies

Douglas N. Honorof and D. H. Whalen
Haskins Laboratories, 300 George Street, Suite 900, New Haven, Connecticut 06511

(Received 7 January 2010; revised 19 July 2010; accepted 18 August 2010)

To identify a speaker's sex, listeners may rely on sex-based differences in average fundamental frequency (F0), but overlap in male and female F0 ranges undermines such judgments. To test accuracy of sex-identification throughout the F0 range, listeners were asked to judge sex based on audio recordings of /ɑ/ spoken on a number of overlapping steady F0s by 10 male and 10 female English speakers. In general, listeners performed above chance (71.6% correct). However, near range extrema, listeners followed an apparent bias toward hearing high F0s as female and low as male; confidence was high when accuracy was high and vice-versa. At mid-range, listeners identified sex fairly accurately but were not very confident in their judgments. In a forced-choice task, vowels close in F0 (but beyond the difference limen) were presented in male-female or female-male pairs. Listeners weakly identified speaker sex (63.3% correct). Identification of the male voice was considerably above chance only when the male had the lower F0 of the pair. Reliance on stereotypes of speaking F0 may bias listeners to hear low F0s as male and high F0s as female, perhaps with a contribution from vocal-tract length information. No strong evidence for a contribution of voice quality obtained.
 [DOI: 10.1121/1.3488347]

PACS number(s): 43.71.Bp, 43.71.An [MSS] Pages: 3095–3104

I. INTRODUCTION

It is possible to identify a speaker's sex from the audio signal alone (e.g., Lass *et al.*, 1976). Early reports suggested that, when gauging the sex of adult speakers, listeners may rely upon resonances of the vocal tract even when judging from voiceless fricatives (Schwartz, 1968; Ingemann, 1968; Schwartz and Rine, 1968). It has also been reported that listeners can reliably identify speaker sex from filter properties alone even when source properties are unusual as with whispered speech, esophageal speech or the use of an artificial larynx (Coleman, 1971; Brown and Feinstein, 1977; Lass *et al.*, 1976; Weinberg and Bennett, 1971) and under conditions of temporal deformation of the signal that do not alter speaker characteristics (Lass *et al.*, 1978). Presumably, sex identification from filter properties would be possible because formant frequency correlates strongly with vocal-tract length (Fant, 1960), or more specifically with the averaged distance between successive formant frequencies, apparently even for macaques (Fitch, 1997), and because vocal tract length in turn correlates with body size (Smith and Patterson, 2005), which correlates with sex. The link between sex and supra-laryngeal vocal-tract length (or more indirectly sex and skull size) emerges in puberty when the trajectory of development diverges for boys whose vocal tracts lengthen more than those of girls, concomitant with changes in the relative sizes of the oral and pharyngeal cavities (Fitch and Giedd, 1999).

Researchers have also reported that listeners appear to attend not just to filter properties, but also, or even primarily, to relative fundamental frequency of vocal fold vibration (*F*0) in identifying the sex of adult speakers (Coleman, 1971; Bachorowski and Owren, 1999; Smith *et al.*, 2005, 2007; Whiteside, 1998), and that average F0 may be especially important to cueing maleness (Coleman, 1976). Hillenbrand and Clark (2009) used resynthesis to manipulate source and filter properties. They found that F0 was somewhat more effective in signaling speaker sex than formants, but that both were necessary for robust identifiability. Presumably, this is possible because F0, like vocal-tract length, correlates strongly with speaker sex. Indeed, among anatomically mature speakers, speaker sex is the greatest predictor of average F0, with males exhibiting an average F0 of about three-quarters of an octave lower than females in citation-form monosyllables (Peterson and Barney, 1952; Hillenbrand *et al.*, 1995; see also van Bezooyen, 1984, on sex, F0 and perception in short phrases). The link between sex and F0 emerges in puberty due to hormonal changes in the male that trigger hypertrophic growth of laryngeal anatomy (Beckford *et al.*, 1985). (For discussions of specific glottal characteristics that correlate with sex and a comparison of auditory versus higher-level processing of sex characteristics of the voice, see Hanson and Chuang, 1999; Mullennix *et al.*, 1995, respectively.)

Although much of the empirical work in this area has relied upon experimental manipulations of average sex-based F0, the real-world task for the listener is more complex. The human voice does not linger at a sex-specific or speaker-specific average F0. Rather, speakers use a wide range of F0s to distinguish words via stress, pitch-accent, lexical tone, level tones, grammatical tone and intonation (Lehiste, 1970; Childs, 2003; Bolinger, 1964; cf. McDonough, 2002, for a possible exception in Navajo where intonation is concerned). Furthermore, in no case is linguistic structure encoded in terms of absolute F0. Rather, F0 level tones and contours shift up and down as a function of the range of F0 values favored by, and to some extent anatomically limited for, the individual speaker (Leather, 1983; Moore and Jongman, 1997).

To complicate matters further, average F0 ranges are less sex-linked than are average F0 values. Honorof and Whalen (2005) report that F0 ranges for 10 men and 10 women overlapped considerably—for some of the speakers, entirely—but that listeners are surprisingly successful at gauging where within an unfamiliar speaker's range the F0 of a particular spoken vowel lies even without sentential context (as in, e.g., Wong and Diehl, 2003). Accounting for the data in Honorof and Whalen (2005) is no simple matter. F0 does not correlate strongly with body size or vocal tract length. The relatively poor correlations between filter properties and F0 suggest that perceptual normalization of F0 location-in-range cannot be based on probabilistic estimation of vocal tract length (Lass and Brown, 1978; van Dommelen and Moxness, 1995; González, 2004; but see Assmann and Nearey, 2007, for a claim that the moderate correlation that does exist matters to listeners). Honorof and Whalen (2005) propose three alternate explanations for how the listener knows where an isolated F0 lies within an unfamiliar speaker's F0 range. Specifically, the investigators speculate that, 1) listeners might simply parse location-in-range out of the signal on the basis of voice quality cues; 2) listeners might normalize F0 within range on the basis of stored population values; or 3) listeners might first identify the sex of the speaker, and then normalize on the basis of sex-based population means.

In order to establish a possible empirical basis that would make the third explanation, above, viable, results of a pair of experiments are reported herein. They aim to addresses whether human listeners can identify the sexes of unfamiliar speakers on the basis of isolated vowels alone *throughout a wide range of F0 values*, where F0 ranges overlap considerably between the sexes. Statistical analyses will test whether accuracy and confidence of sex identification co-vary with location of pitch within the F0 range. The hypothesis to be tested is that listeners will be sensitive to speaker sex even when presented only with the isolated vowels of unfamiliar speakers, and that listener sensitivity to sex will be apparent at all locations within the speaker's F0 range.

II. EXPERIMENT 1: NONPAIRED SCALAR IDENTIFICATION AND CONFIDENCE

Such potential vocal correlates of speaker sex as spectral envelope should be relatively independent of F0 and could thus be presumed to be accessible throughout the range. However, because previous tests used sentences with time-varying intonation contours (e.g., Lass *et al.*, 1978; Hillenbrand and Clark, 2009), it is unclear whether speaker sex information is available at more extreme F0s. Furthermore, language-specific cues to sex may have been confounded with purely biological factors in the materials in those studies. Therefore, Experiment 1 was designed to test how well listeners could extract information about a speaker's sex from isolated vowels throughout the range.

A. Materials

1. Speakers

The speakers and stimuli were those of Honorof and Whalen (2005) and are described in greater detail therein. To summarize, 10 male and 10 female native speakers of North American English produced the stimuli used here. Their ages ranged from 20 to 78 (*M* 33.25, *SD* 14.74). Although the 78-year-old male outlier was 24 years older than the oldest female and 37 years older than the next oldest male, mean age and standard deviation in age were similar for the two sexes. Among the women, mean age was 32.5 (*SD* 13.17). Among the men, mean age was 34.0 (*SD* 16.90). A one-way ANOVA (SPSS 16 for Mac OS X, SPSS, Inc., Chicago, Illinois) found no significant difference in speaker age between the two sexes [$F(1,19)=0.49$, $p<0.828$].

Prior to participating, all speakers confirmed that they had no training or experience in vocal solo performance, that they had not smoked for at least five years, that they were not suffering from congestion, coughing, post-nasal drip, or fatigue, that they did not suffer from severe or chronic respiratory allergies, and that they had never been diagnosed with a clinical condition that might interfere with speech production such as a communication disorder, voice pathology, or a neuromotor impairment. On the day the stimuli were collected, each speaker passed a binaural audiometric screening performed in a sound-attenuated chamber using a Beltone 110 audiometer at pulsed frequencies between 0.125 and 1 kHz presented at 25 dB HL or lower–well below the average amplitude of conversational speech.

2. Materials

For purposes of stimulus creation, speaker-specific spoken F0 ranges were elicited via spoken glissandi with falsetto excluded. Each participant then spoke the vowel /ɑ/, each time on a distinct F0 sustained for four to five seconds. Production of each vowel was prompted by loudspeaker presentation of a 300 ms sinusoid on at least 8 musical pitches, always C, E, G or A ("tuned" relative to Concert A). Across speakers, the lowest frequency attempted was 49 Hz and the highest 524 Hz. Real-time visual feedback on pitch-matching was provided to the speakers via a Kay Visipitch II. The vowels were recorded with an Earthworks QTC1 microphone and digitized in real-time at 44.1 kHz, 16-bits. For the listening test, eight tokens for each speaker were selected so that they were spread roughly evenly throughout the speaker's F0 range. Each vowel was trimmed to the first 500 ms that contained no noticeable F0 or intensity excursions. A linear amplitude ramp (100 ms) was applied at the beginning and end of the vowel segment selected. For one female speaker, the highest sustained non-falsetto F0 values lay 3% and 14% above her falsetto break as found in her glissandos, and 5% above for another. At the low end, all vowels selected lay within the spoken F0 range. Finally, as described in Honorof and Whalen (2005), location-within-range as a percentage was calculated for each F0 achieved.

3. Listeners

Six men and six women participated in the listening tasks. Their ages ranged from 20 to 51. Prior to participation, they were asked to confirm that they were non-disordered native speakers of English who had never been fluent in a second language. All also indicated their ability to use a computer monitor and mouse. On the test day, each listener passed a binaural screening via a Beltone 110 audiometer in a sound booth. Tests confirmed that the listeners were able to hear audiometric pulsed frequencies of 0.125, 0.5, 1, 2, 4 and 8 kHz at 30 dB HL or lower in both ears. In a debriefing interview, all were asked questions relevant to their participation. All 12 respondents indicated that they had never suffered from a speech, hearing, language or communication disorder, with the exception of one listener who reported having stuttered due to stress between the ages of 8 and 10. That respondent reported that she out-grew the stutter without clinical intervention. All 12 indicated that they were raised by primary caregivers with no known history of speech, hearing, language or communication disorders.

4. Procedures

Listeners were seated in a quiet room in front of a MacBook Pro running MATLAB R2007b (The MathWorks, Inc., Natick, MA) and asked to read a set of instructions (see Appendix A). They were fitted with a pair of HD 280 Pro closed, circumaural headphones (Sennheiser Electronic Corporation, Old Lyme, CT). In the first phase of presentation, listeners heard a single token from a single speaker and were presented with a MATLAB Graphical User Interface (henceforth, *GUI window*) that posed the question, "Is the speaker…?" with response buttons aligned on a vertical axis. There were two presentation conditions. In one, all male response buttons appeared beneath all female response buttons. In that condition, the text label at the top of the GUI window read "Female." That text headed a column of buttons themselves labeled "Female for sure" "Probably female" "Possibly female" and "Guessing female." Below the last female button, there was a long horizontal dash, then four more buttons labeled "Guessing male" "Possibly male" "Probably male" and "Male for sure" and beneath a text label "Male." A "Repeat" button was also provided to allow confirmation of difficult judgments. Listeners were discouraged from listening more than twice. No data on number of times repeated were recorded, but no subject took noticeably longer than the others to complete the task. In the other presentation condition, the GUI window was the same, but the order was reversed with the corresponding female text and response buttons appearing beneath the male text and response buttons. The inversion of the sex-association of the response buttons was intended to avoid associating male F0s with low position on the screen.

a. Presentation. All eight tokens from all 20 speakers were randomized together in two blocks and presented diotically at a comfortable level set by the participant during a presentation of 10 practice trials. (The practice trials were produced by the first author and consisted of stimuli similar to those of the experimental trials.) Listeners indicated their identifications via mouse click.

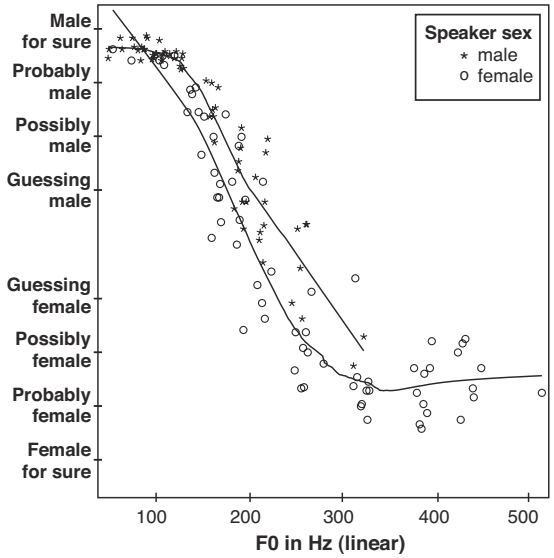

FIG. 1. Experiment 1. Estimated marginal means of confidence ratings for each unique stimulus pooled across listeners against F0 (linear). Separate fit curves for vowels produced by men (dashed curve) and women (solid curve) are provided. Curves shape at the low end for males and high end for females suggest "ceiling" effects. At the mid-range, the male and female curves are not very different, but the male curve suggests more confidently male responses at a given frequency with male voices (i.e., a higher curve), and more confidently female responses at a given frequency (i.e., a lower curve). However, location of the marginal means with respect to the most extreme labels on the vertical axis suggests that listeners were not as certain that high F0s were produced by females as they were that low F0s were produced by males, even though male and female F0s were both present in the low end of the distribution, but not the high end.

B. Experiment 1: Results and discussion

Any male response to a male stimulus regardless of confidence was coded as correct; any female response to a female stimulus regardless of confidence was coded as correct. The mean accuracy across all 12 listeners and stimuli was 71.61%, with chance at 50%. An independent-samples t-test was run to compare accuracy in male-response-button-low versus female-response-button-low conditions. There was no significant difference in the scores for male buttons low ($M=0.72$, $SD=0.447$) or female buttons low ($M=0.71$, $SD=0.455$); $t(3,838)=1.145$, $p=0.252$ with equal variances not assumed ($F=5.246$, Levene's $p<0.023$). Consequently, button order was excluded from further analysis.

Figure 1 shows confidence in sex judgment of each stimulus averaged across the 12 listeners. An estimated marginal mean in the top half of the figure indicates that the speaker was judged to be within the male response scale. An estimated marginal mean in the bottom half of the figure indicates that the speaker was judged to be within the female response scale.

Curves were fitted to the male and female speakers separately by local regression (LOESS fit method) with a biweight kernel (50% of points). Local regression was used because its iterative re-weighting is especially robust against outliers. Evident changes in the curve at both ends of the

distribution reflect "ceiling" effects which indicate that low F0s were heard as male and high F0s as female irrespective of the actual sex of the talker. In the mid-frequency region, male responses to male stimuli were pushed toward the male end of the confidence scale and female responses to female stimuli were pushed toward the female end of the scale. Thus, the fit lines for the two sexes are distinct in this mid-frequency region which indicates a degree of correct perception of sex.

In order to quantify local F0 effects on accuracy of sex judgment and to explore possible group effects of talker sex and subgroups of listeners, a multilevel mixed-effects model was fitted to the data via hierarchical linear modeling (HLM). Hierarchical modeling was possible because each listener heard all the male and all the female pitches. Although pitch was nested within speaker sex, the exact pitches differed from voice to voice so pitches were simply ranked ordinally from low to high for the males and low to high for the females, 80 ranks for each sex. In cases in which pitches were identical for two or more speakers, local ranking was arbitrary.

In the modeling, speaker sex and F0 rank were treated as fixed factors. In order to adjust for possible sampling variation of any F0 ranking effect due to different speaker sexes and unanticipated differences among listeners, F0 rank was treated as a random effect as well.

Under Type III testing, there were significant results for the intercept-only model $[F(1,1839)=7453.075,\ p<0.001]$, speaker sex $[F(1,1839)=173.848,\ p<0.001]$ and F0 rank $[F(79,1839)=3.723,\ p<0.001]$. Estimates of covariance parameters confirm that speaker sex was judged less accurately at the two ends of the distribution. The HLM failed to produce statistical evidence for *listener-based* response differences to sexes and levels; the estimate is small and the Wald Z not significant for variation of F0 rank due to speaker sex by listener.

In order to explore the portion of the data that appeared to be free of ceiling effects, namely, the mid-frequency region, and in order to more fully explore confidence and accuracy, data were binned as follows. All data 150 Hz and below were grouped together. All data 271 Hz and above were grouped together. The remaining data were grouped into four bins each spanning a range of 30 Hz. The binned response data appear in Fig. 2. Accuracy as a percent correct accompanies each data point in parentheses.

Conditions in which mean accuracy was below chance are indicated with an 'X'. There were four such conditions: females in the two lowest bins and males in the two highest bins. Accuracy was below chance in exactly those cases in which F0 stereotypes are violated: from 4% to 31% correct for females in the lowest two sixths of the global frequency range and from 15% to 41% correct for males in the highest sixth of the global frequency range. In other words, the male–female pattern is symmetrical with respect to accuracy and frequency bin. For the misidentified females (that is, with two bins pooled), confidence was between "Guessing male" and "Probably male" ($N=648$). Listeners were even less confidently wrong for the misidentified males (again, with two bins pooled); confidence in wrong responses to

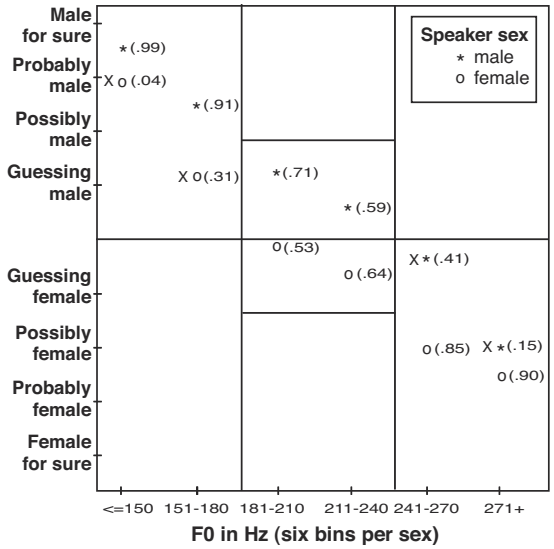

FIG. 2. Experiment 1. Confidence ratings. Values closer to the horizontal line that spans the figure reflect relatively low confidence and vice-versa. Accuracy is given in parentheses for each point. The high and low bins are based on the inflection points of the curve in Fig. 1. The other four bins evenly divide the remaining F0 values.

male items ranged around "Guessing female" ($N=192$). To break inaccurate responses down further, confidence was poor in the second and fifth bins, but better ("possibly" or nearly "probably") in the most extreme F0 bins.

At the extrema, accuracy was very high in exactly those cases in which F0 stereotypes were upheld: from 90% correct to 99% correct. In the middle two frequency bins, overall accuracy is always above chance without respect for stereotype conformity, but lower than for the correctly judged bins with more extreme F0 values. Stereotype-based expectations may provide a plausible explanation for the pattern of response bias at the extrema. That is, it may be that listeners expected males to be more prone to very low F0 and, to a lesser extent, females to be more prone to higher F0, with such stereotypes perhaps constructed as proposed in Eckert (2008).

Strikingly, confidence ratings paralleled accuracy ratings for accurately judged bins. Highly accurate bins were judged with the greatest confidence and less accurate bins (mid F0 range) with less confidence. On average, in all bins, when listeners were wrong, they were less confidently wrong than they were confidently right. This pattern reflects a tendency for confidence to push toward the correct sex judgment whether the mean judgments were correct or not. It seems that listeners know at some level that they are wrong when relying so heavily on F0 in the face of conflicting formant cues, but are more certain that they are right when F0 and formant cues converge. Listeners indicate even lower certainty yet when F0 is expected to provide little benefit, that is, the middle frequency ranges. In those frequency bins, listeners still have formant information and perform above chance, but, in the absence of helpful information from F0, report that they are guessing or nearly so, even though they are not. Surprisingly, the most ambiguous stimuli with respect to F0 are judged with even lower confidence than higher and lower frequency stimuli judged wrongly. Listen-

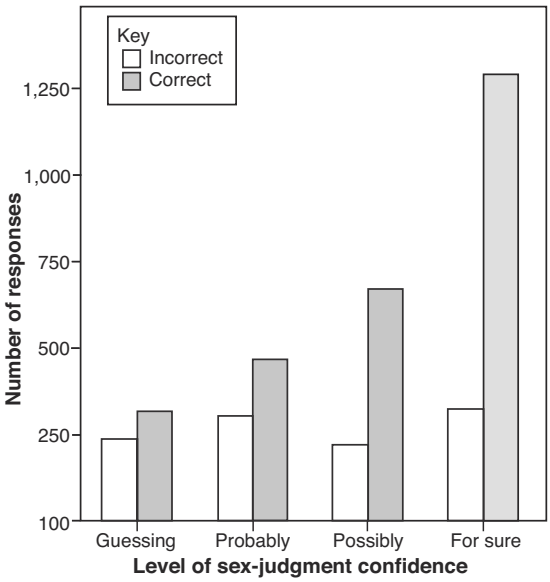

FIG. 3. Experiment 1. Confidence broken down by accuracy of sex judgment.

ers trust F0, but they are most confident when F0, formant structure and any other cues present in the stimuli tell a consistent story.

Another pattern emerges from the confidence ratings. If one compares confidence for bins symmetrically (that is, bin 1 male with bin 6 female, bin 2 male, with bin 5 female, etc.), confidence is higher for the males in every part of the global F0 range, suggesting that the sex of male voices may be easier to perceive, a finding consistent with that of Owren *et al.* (2007).

When the data are split according to accuracy of response and sorted according to confidence, a pattern of association emerges (see Fig. 3). A cross-tabulated symmetric correlation, specifically a measure of monotonicity for which no directionality was assumed, tested the extent to which degree of confidence (right or wrong) predicts overall accuracy. A significant ($p<0.001$) Goodman and Kruskal's gamma value of 0.325 indicates that there is a deterministic and concordant association between the variables. That is, nearly a third of the variation in accuracy is accounted for by variation in confidence, and the two variables change in the same direction.

III. EXPERIMENT 2: PAIRED FORCED-CHOICE IDENTIFICATION

Experiment 2 was designed to reduce the listeners' reliance on absolute F0 in judging sex. Stimuli were paired–one F0 from a male and one from a female. Within a pair, F0 difference was kept as close as possible. The differences were always above and generally just above the difference limen (Wier *et al.*, 1977).

A. Methods

1. Procedure

Immediately following the first experiment and a brief unrelated experiment, the 12 listeners of Experiment 1 heard pairs of stimuli which they were told contained a male and a female voice (see Appendix B). On each trial, a selected token from a single speaker was followed after 750 ms by a selected token from a speaker of the opposite sex. Listeners were prompted by GUI window to report, "Which speaker is male?" with two response buttons aligned on a vertical axis and labeled "1" and "2." Judgments were entered by mouse-click. A "Repeat" button was provided, though listeners were discouraged from listening more than twice. No data were collected on the number of repeated items, but no listener took noticeably longer than the others at the task.

Each listener heard each of 20 vowel pairs four times. Across all pairings, half the time the male vowel was lower in F0 than the female vowel, and half the time higher. In two out of four blocks (randomized once by block), the first of the two vowels presented was spoken by the male pair member half the time. In an additional two blocks (re-randomized), this pattern was reversed. Thus in half the trials across blocks, the correct answer to "Which speaker is male?" was "1" and vice-versa.

F0 distance between pair members ranged from 6 to 14 Hz, with an additional two pairings with male and female members 107 and 109 Hz apart. In all cases, F0 distance between pair members was easily above the difference limen (Wier *et al.*, 1977). All tokens were selected from among the stimuli of Experiment 1 so as to use all 20 voices at least once and no single token from a single speaker more than once, and also to minimize distance in F0 between pair members, and so as to disperse the pairings throughout the pooled population F0 range insofar as possible. All listeners were presented with the same materials in the same pseudorandom order. Pairings were presented diotically at the same level as in Experiment 1. There were no practice trials prior to the 80 test trials.

B. Experiment 2: Results and discussion

When the female vowel had a lower F0 than the male vowel with which it was paired, mean accuracy in sex identification pooled across listeners was barely above chance (55.6% correct, *SD* 9.09, $N=480$). When the male member of the pair was lower in F0, however, accuracy was higher (71.0% correct, *SD* 13.63, $N=480$). Here overall accuracy declined to 63.3% correct (*SD* 13.91, $N=960$) from the 71.6% correct of Experiment 1. Apparently, building in the assurance that one of the speakers was male and one female did not make the task easier for the listeners as anticipated; sex-identification was not especially strong.

When pairings were split into low and relatively high halves of the population F0 range (10 pairings in each half), a marginal pattern emerged. When the lower F0 had been produced by a female speaker, sex was identified slightly more accurately when the F0s paired were above 200 Hz (58.8% correct, *SD* 9.28, $N=240$), but performance was even closer to chance below 200 Hz (52.5% correct, *SD* 7.74, $N=240$). Apparently, if the female was lower than the male and low with respect to the population range, listeners resorted to a guessing strategy, perhaps due to a bias against low F0 being produced by women or because any sex-information

that might be present at mid- and high-range was simply missing when the female speakers were pushed near the low end of their F0 ranges. This reflects the pattern of Experiment 1. In Experiment 2, symmetric results obtained when the male vowels were produced with a lower F0 than the female vowels. Specifically, the sex of the male speakers was identified with greater accuracy when the F0 judged was below 200 Hz (75.0% correct, *SD* 11.66, $N=240$), but only with 67.1% accuracy when the F0 judged was above 200 Hz (*SD* 14.31, $N=240$).

One might speculate that the task of identifying speaker sex would be made easier by the inclusion of two pairings in which the distance between the F0s was relatively large, especially if the male F0 is the lower of the two. However, the percent correct did not differ much when the two pairings with large F0 differences were removed from the calculation. With the low-F0-large-distance pair removed, there was indeed a slight reduction in accuracy when males were lower than females at the lower half of the population F0 range (down nearly 5% to 70.3% correct, *SD* 7.73, $N=192$). Reduced performance is expected here because the low-in-range-large-distance pair had a distance of over one octave (male 76 Hz, female, 183 Hz), which would have been easy for listeners to tell apart if they were using F0 as a gauge. Removing the high-F0-large-distance pair resulted in only a small increase to overall judgment accuracy above 200 Hz (from 62.92% correct to 65.28% correct, *SD* 11.17, $N=432$). In that case, the F0 distance between members of the pair that was removed was even greater (109 Hz), but nowhere near a doubling of pitch in that part of the F0 range. Overall it would appear that relatively weak performance on the paired, forced-choice task was not an artifact of inclusion of outliers in F0 distance. Because excluding the two obvious F0-distance outliers did not make much of a difference in the descriptive statistics, all tokens were included in subsequent inferential statistical analyses.

A univariate ANOVA was run to assess the effects of two fixed factors: F0 order (whether the male or female pair member had the lower F0) and F0 range bin (whether the F0 of both pair members fell below or above 200 Hz) on a dependent measure of percent correct for each pairing across listeners and blocks. Thus the analysis comprised a two-by-two between-subjects factorial design.

The ANOVA revealed no significant main effect for F0 range bin $[F(1,956)=1.371, ns]$. However, a significant main effect for F0 order $[F(1,956)=468.73,\ p<0.001]$ obtained, and a significant interaction between these factors was observed $[F(1,956)=98.96,\ p<0.001]$. An associated multiple linear regression gave an R^2 of 0.37 indicating that the two factors were strongly related (see Fig. 4).

As reflected in the pattern of means described above, sex was judged more accurately when the male voice was lower than the female voice, especially in the low F0 range bin. A one-way ANOVA revealed a significant difference among the four means, with a significant overall one-way effect $[F(3,956)=189.69,\ p<0.001]$. Post-hoc Tukey tests revealed that the differences in means reflected in the interaction are significant at the $p<0.001$ level for all comparisons ($N=240$ for each cell).

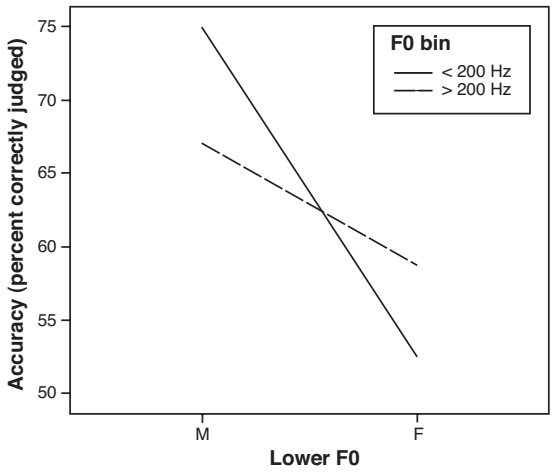

FIG. 4. Experiment 2. Significant interaction between whole-pair F0 range bin (above or below 200 Hz) and whether or not the male F0 was the lower of the two; 12 listeners pooled. The dependent variable is accuracy calculated as a global percent correct for each pair.

The relative difficulty listeners had in correctly identifying the sex of female voices in the low end of the range (when the paired male voice was higher) could indicate that there was something unusual about the voice quality of only the women at the low end of the range. Therefore, automatic jitter and shimmer analysis was performed on all pair members using the voice report functions in Praat 5.1.33 (Boersma, 2001). To exclude variability in glottal pulse characteristics potentially associated with the onset and offset of the vowel, glottal pulse characteristics were inferred on the basis of only the central 300 ms of each vowel. However, in one case (a male 65 Hz vowel), the automatic phase-point-picking algorithm overidentified pulses at higher harmonics during eight cycles roughly centered around the temporal midpoint of the vowel. Extraneous pulse labels were removed by hand, and the remaining pulses regularized with respect to phase-point. A voice report analysis was run on all vowels with the setting (40–400 Hz) held constant across all talkers in order to allow the full range of F0s to emerge. Four automatic measures were obtained: absolute jitter (average difference between periods; threshold 83.2 μs), F0-normalized jitter (average difference between periods divided by the average period; threshold 1.04%), shimmer (average difference in amplitude between periods divided by the average amplitude; threshold 3.81%) and an SPL-adjusted shimmer (average base-10 logarithm of the difference in amplitude between periods multiplied by 20; threshold 0.35 dB). See Brockmann *et al.* (2008), for a discussion of the technique. Only five vowels had above threshold levels of jitter according to both measures. They were the two lowest male F0s and the three lowest female F0s. In addition, the next lowest male and next lowest female F0 exceeded the jitter threshold for absolute jitter only. No other tokens exceeded jitter thresholds. All vowels that exceeded jitter thresholds also exceeded both shimmer thresholds. Figure 5 shows the sex judgment accuracy for each F0 pair.

Issues of F0 range extrema and stereotype aside, listeners were relatively successful in accurately judging sex from the voice, especially at mid-range. Given that modulated

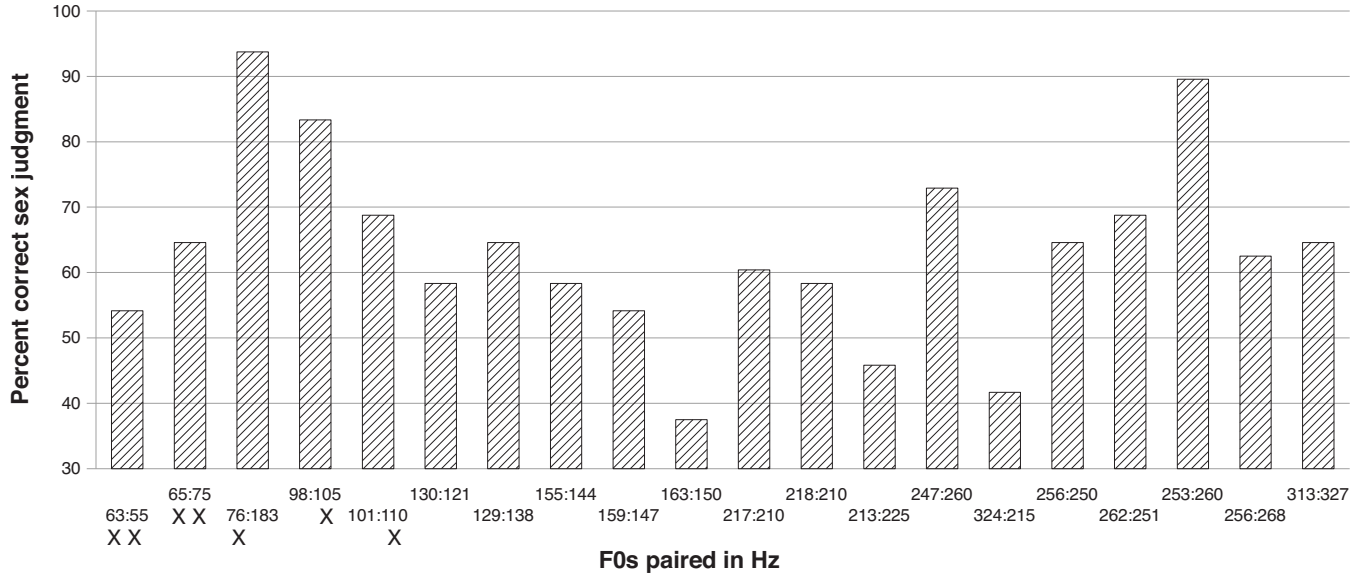

FIG. 5. Experiment 2. Sex judgment accuracy by pairing. The F0s for each pair member are given in the horizontal axis labels. In each case, the left-most number is the male F0 irrespective of order of presentation and irrespective of which number is lower. An area of relatively inaccurately judged female-lower pairs can be seen, especially in the low end of the population F0 range and a symmetrical clustering of relatively accurately judged male-lower pairs can be seen throughout. The seven vowels with over-threshold levels of both jitter and shimmer are marked with an 'X', and lie exclusively at the low end of the F0 axis. Accuracy ranged from 54% to 83% correct with no obvious pattern that would suggest a correlation between jitter/shimmer and accuracy in sex judgment.

contours in F0 and intensity have largely been excluded from the stimuli, it remains to determine exactly where the sex-cueing information about speaker sex lies in the present stimuli, even if sex cues are overridden by F0 expectations at the extrema.

Although the foregoing results of the jitter and shimmer analysis do not entirely rule out other possible sex differences in glottal characteristics, an at least equally plausible cue to sex could lie in formant patterns with males perhaps having a longer vocal tract and lower formants. (Again, see Fant, 1960; also Ives *et al.*, 2005.) Therefore, we tracked the first three formants from the all items included in either experiment (thus all stimuli from Experiment 1) via LPC analysis using the Burg method in Praat. Parameters, including amount of pre-emphasis, were adjusted on a case-by-case basis until tracked formant values appeared plausible when compared with wide-band spectrograms also generated on a case-by-case basis. Unfortunately, analysis was not always a simple matter. When F0 was high, F1 and F2 sometimes appeared to merge in the wideband spectrogram, forcing us to rely solely on LPC tracks. Furthermore, F3 was often weak presumably due to the addition of nasal antiresonances that typically accompany open vowels. F3 was often especially weak for some of the highest F0s which would have wide harmonic spacing. In all cases, multiple passes were made until values were estimated that appeared plausible with visual reference to the spectrogram. Results are plotted in Fig. 6.

F1 and F2 differed by sex with males having the lower mean formants, while F3 did not reach significance (see Table I). Thus there was some information about sex that was available in the formant pattern. F0 itself was significantly higher for the women and positively correlated with F1 and F2 ($r=0.372$ and 0.524, respectively; $p<0.001$).

IV. CONCLUSIONS

In two experiments, the listener's ability to determine speaker sex from an isolated vowel was assessed. Identification across F0 range was significantly accurate in the middle frequencies in Experiment 1 with accuracy apparently modulated according to a stereotype-consistent response bias. A multilevel mixed-effects model confirmed that sex was judged either significantly more or significantly less accurately at the extrema depending on the relationship between the F0 and the correct answer. At low frequencies, males were more accurately identified but females were not well-identified. At high frequencies, females were more accurately identified while males were misidentified as female. Thus the results failed to fully confirm the prediction that listeners

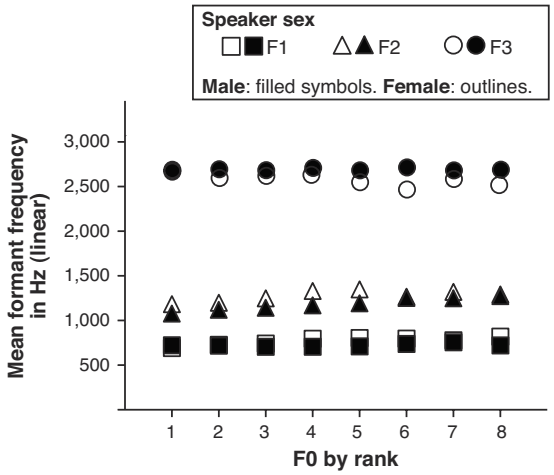

FIG. 6. Experiments 1 and 2. Mean LPC-tracked formant values for the first three formants coded by speaker sex. Averaging is across speakers within F0 rank with low F0 on the left.

TABLE I. Univariate ANOVAs: Effect of speaker sex on formant frequency.

DV	Female (*M*)	Male (*M*)	F(1, 18)	*p*
F0	256	153	90.31	<0.001
F1	769	711	4.57	0.046
F2	1273	1174	7.87	0.012
F3	2569	2681	3.32	0.085

would be able to identify sex of an unfamiliar talker from isolated vowels throughout his or her range. Gamma correlations confirmed that confidence predicts accuracy significantly. Confidence pushes toward the correct sex judgment regardless of whether the mean judgments were right or wrong, which suggests that listeners may extract sex information but may nevertheless sometimes reject it when F0 is either misleading or simply uninformative. In the second study, lowered uncertainty was predicted for a forced-choice paired presentation with a male and female talker presented on each trial, but accuracy did not increase as predicted. Listeners were especially inaccurate when the female voice was the lower of the two at low frequencies. Acoustic measures of voice quality for the larger data set of Experiment 1 (of which the data of Experiment 2 were a subset) did not indicate any voice quality difference that would explain the results. The formants, however, differed by sex, and therefore might have been expected to support more accuracy than was obtained, especially at the high end of the F0 ranges where there was less overlap between the sexes. Although vocal tract lengths might have contributed to sex-judgment, that contribution may have been offset by other factors such as whether the male F0 was the lower of the two as would be consistent with stereotype, and then so in a bin-specific pattern.

Earlier studies have reported higher overall accuracy in identifying sex from the voice than was evident in either of the present experiments. By virtue of having presented stimuli at a range of F0 values, the present experiments were able to explore sex identification throughout a population-defined F0 range independent of language structure (intonation, etc.). The present stimuli incorporated individual ranges that overlapped considerably, especially at the low end. In contrast, studies that rely on one average male F0 value and one average female F0 value, for instance, effectively reduce the potential for error to a virtual "flip of a coin," making it relatively easy for listeners to judge sex accurately. When those two values differ by an octave, a high degree of accuracy indicates only that one can hear octaves and associate them by stereotype with male and female sex categories. In the present pair of experiments, the listener's task was more conservative; it was not possible to familiarize oneself quickly with the F0 values one might hear as there were too many present in the stimuli. Consequently, in neither of the present tasks did listeners perform as well as expected based on earlier studies. However, the inclusion of a wide range of values allowed for binning of values and analysis of different parts of the F0 range, which revealed a number of coherent patterns.

Specifically, in Experiment 1, on average, listeners judged speaker sex accurately when low F0s were produced by males and high F0s were produced by females, that is, when the sex-F0 relation reinforced sex stereotypes for adult F0–stereotypes. In the population of speakers in the present study, accurate judgments reflect male and female differences in average or ceiling F0 rather than basal F0 or F0 ranges. Confidence was relatively high. Conversely, on average, men high in their ranges were often misidentified as women, and women low in their ranges were often misidentified as men, but in these conditions in which stereotype was contravened, confidence was slightly to considerably lower. This pattern of lowered confidence for inaccurately judged items, may suggest that at some level, listeners really did know that they were not reporting sex accurately, but were biased in their reporting. It was as though listeners could identify sex correctly (perhaps from vocal tract length as reflected in the formant patterns), but were unwilling to admit that women could speak vowels on very low notes or men on very high notes. Listeners were not simply guessing. In the middle of the global (that is, cross-speaker) F0 range, however, listeners were above chance at sex identification, but not nearly as far above chance as they were for the correctly judged extreme F0s. Percent correct in the middle two bins ranged from only 53% to 71%—somewhat better than guessing. Here, however, confidence was as low as in the inaccurately judged conditions or even lower. Taken as a whole, results suggest that the very superior and the very inferior performance at the high and low ends of the range reflect the listener's usual "fair" ability to identify sex from isolated vowels of unfamiliar speakers (as seen clearly in the middle of the F0 range) plus a bump up or step down introduced by a bias in what listeners are willing to report.

Another pattern emerges in Experiment 1. Namely, in all bins judged accurately, men are judged with greater confidence than women in the same bin. Furthermore, if one adopts the assumption of symmetrical bias and compares bin 1 with bin 6, 2 with 5, 3 with 4 for the men and the inverse for the women, the men are also judged with greater confidence for each comparison (cf. Fig. 2). Better performance at sex identification of male voices is consistent with the findings of Owren *et al.* (2007) who speculate that a perceptual advantage for male sounds in sex discrimination follows from the fact that dimorphism in vocal anatomy is asymmetrical from a biological or evolutionary perspective; boys diverge from a developmental trajectory shared with girls, not the other way around. Adult male voices are more different from the voices of boys and girls than are adult female voices; men's voices stand out with respect to the population as a whole.

In an effort to force listeners to judge sex from whatever information might be available beyond F0, we paired male and female F0s that were close in value and presented them in a forced-choice task (Experiment 2). Listeners were told that one of the paired tokens was spoken by a man and the other by a woman; cognitive load was very low. The pairing of items in this way should have undermined reporting bias. The instruction to treat one pair member as male and the other as female implied that not all the low pairs could be

male and all the high pairs female. Nevertheless, performance was only somewhat above chance, though better when the male was the lower of the two, especially below 200 Hz. These results are consistent with the following interpretation. Listeners again attended to whatever they attended to in the non-paired experiment, and they still tended to report the lower member of the pair as male especially for low F0 pairs, but were reluctant to do so to the same extent, leading to fewer bias-enhanced false positives and bias-diminished false negatives. A guessing strategy would have resulted in lower scores yet.

Post-test debriefing revealed that some listeners found at least some of the stimuli remarkable in ways that are not always easy to map into scientific terminology straightforwardly. The present stimuli were not, however, entirely artificial. They were naturally produced–not synthesized–and in most respects (except in the domain of prosody) similar to isolated vowels spoken outside the laboratory setting. The vowel quality itself was unremarkable. Certainly isolated monophthongs exist in realm of English expressives ('uh', 'ah', etc.). However, by design, the stimuli themselves were unusual. First, a few vowels contained jitter and shimmer, especially those vowels that were extremely low in F0, including those F0s that were "unexpectedly" low for female voices. Jitter and shimmer are not rare, but may indicate that our speakers were pushed to extrema that they may or may not normally use in their speech, especially the women. The more obvious prosodic peculiarity of our stimuli lay in their monotonicity; they were spoken (not sung) on artificially steady F0s at a relatively steady intensity. The monotone imposed was intended to make nonlinguistic sex information stand out, which should have made sex identification easier than in everyday life, at least insofar as sex information can be attributed to biology as opposed to learning. Certainly, sociolinguistic investigation (e.g., McConnell-Ginet, 1978) has suggested a number of learned, sex-specific intonational behaviors involving both F0 and intensity. Such behaviors were absent here, as were any possible sex-correlated grammatical constructions (more a consideration in some languages than others) or sex-correlated word choice. Whatever intrinsic cues to sex remained in the absence of prosody (e.g., vocal tract length information available especially in the formants) were nevertheless not sufficient to cue reliable sex judgments in all cases. In everyday life, misidentification of the sex of a stranger over the telephone could lead to socially awkward situations to say the least. The present results suggest that if real-world listeners are indeed able to avoid embarrassment by perceiving sex from the voice accurately throughout the voice range, they must do so on the basis of prosody which is absent from our stimuli.

In the study for which the present stimuli were initially collected, Honorof and Whalen (2005) asked whether listeners could rank relative location-within-range of a sustained F0 on a vowel without the opportunity to train up on the voice of the speaker. The investigators observed that, had listeners "relied on a first-pass discrimination of sex to provide two templates for comparison, they would have misranked far more (p. 2199)." Here listeners were explicitly asked to judge sex, which may have biased their attention to certain aspects of the signal. Success was limited. In the only condition in which performance was near ceiling—males in the low parts of their ranges—it was near ceiling due, apparently, to reporting bias. If speaker sex is misidentified in roughly one-third to one-quarter of the cases, determination of sex as a first-pass strategy at F0 normalization would not be an ideal strategy for location of pitch within range.

Because sex might be inferred from a number of acoustic parameters of vowels, the outcome here, though clear, requires a nuanced interpretation. If listeners were to guess at sex on the basis of population norms, they might rely on all of the information present, some of which may be at odds near the extrema within the present stimulus set. Honorof and Whalen (2005) report that the ranges for the stimuli used here contain more gaps for the females, who "seem to have had a difficult time sustaining low F0s (p. 2199)"—frequencies that those individuals were shown to be able to achieve on spoken register glissandos. Furthermore, at the high end where there was less overlap between the sexes, the highest register breaks in the data were achieved by the women and there were many high-end gaps in the sustained F0s spoken by the men.

However appealing a theory of voice quality effects may have seemed despite the absence of evidence that voice quality is a reliable indictor of sex, primary reliance on F0 and secondary reliance on formants provides the best explanation for the overall results here. A follow up study varying F0, intensity, glottal pulse characteristics and vocal tract characteristics with synthetic voices might help disambiguate results further. An analogous natural speech study would be ideal, though voice quality might prove difficult if not impossible for speakers to control consistently alongside intensity and F0. The present results, however, clearly indicate that listeners are not entirely sensitive to whatever intrinsic information about speaker sex may be inherent in isolated vowels, and that their sensitivity to that information does not entirely override reporting bias.

ACKNOWLEDGMENTS

The authors are grateful to Carol Gracco for allowing us the use of her Visipitch. We thank Mitchell S. Sommers, Carol A. Fowler, Laura Koenig and two anonymous reviewers for helpful comments. This work was supported by NIH Grant No. DC-002717 to Haskins Laboratories.

APPENDIX A: EXPERIMENT 1 PARTICIPANT INSTRUCTIONS

You will hear a number of different people saying the vowel "ah." They will be using different parts of their speaking range. Use the mouse to click on a button to let us know whether you think the speaker is male or female, and how sure you are. If you need to repeat a vowel before deciding on whether the speaker is male or female, just hit the Repeat button. You can repeat each item as many times as needed, but, generally, we would prefer that you base your decision on one or two listenings. In some cases, you may not be sure. If you are unsure about an item, just guess and move on to the next one.

You will run through one batch of vowels, perform a slightly different task, and run through another batch. You control the timing, but it should take considerably less than an hour.

If you need to take a rest before or after the scheduled half-way-point break, just stop pressing buttons, and then hit Repeat when you are ready to begin again.

Please ask the experimenter any questions you have now.

Thanks!

APPENDIX B: EXPERIMENT 2 PARTICIPANT INSTRUCTIONS

You will hear two people saying, "ah," one after the other. Click with the mouse to tell us which of the two is male. If you are not sure, guess.

You control the timing, but it should not take more than 10 min.

Assmann, P. F., and Nearey, T. M. (**2007**). "Relationship between fundamental and formant frequencies in voice preference," J. Acoust. Soc. Am. **122**, EL35–EL43.

Bachorowski, J. A., and Owren, M. J. (**1999**). "Acoustic correlates of talker sex and individual talker 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., Schaid, D., and Schanbacher, B. (**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**, 341–345.

Bolinger, D. L. (**1964**). "Around the edge of language: Intonation," Harv. Educ. Rev. **34**, 282–296.

Brockmann, M., Storck, C., Carding, P. N., and Drinnan, M. J. (**2008**). "Voice loudness and gender effects on jitter and shimmer in healthy adults," J. Speech Lang. Hear. Res. **51**, 1152–1160.

Brown, W., Jr., and Feinstein, S. (**1977**). "Speaker sex identification utilizing a constant laryngeal source," Folia Phoniatr (Basel) **29**, 240–248.

Childs, G. T. (**2003**). *An Introduction to African Languages* (John Benjamins, Amsterdam), p. 78.

Coleman, R. O. (**1971**). "Male and female voice quality and its relationship to vowel formant frequencies," J. Speech Hear. Res. **14**, 565–577.

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.

Eckert, P. (**2008**). "Variation and the indexical field," J. Sociolinguist. **12**, 453–476.

Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, The Hague, The Netherlands), p. 242.

Fitch, W. T. (**1997**). "Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques," J. Acoust. Soc. Am. **102**, 1213–1222.

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.

Hanson, H. M., and Chuang, E. S. (**1999**). "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," J. Acoust. Soc. Am. **106**, 1064–1077.

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

Hillenbrand, J. M., and Clark, M. J. (**2009**). "The role of f0 and formant frequencies in distinguishing the voices of men and women," Percept. Psychophys. **71**, 1150–1166.

Honorof, D. N., and Whalen, D. H. (**2005**). "Perception of pitch location within a speaker's F0 range," J. Acoust. Soc. Am. **117**, 2193–2200.

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.

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.

Lass, N. J., Mertz, P. J., and Kimmel, K. L. (**1978**). "The effect of temporal speech alterations on speaker race and sex identifications," Lang Speech **21**, 279–290.

Leather, J. (**1983**). "Speaker normalization in perception of lexical tone," J. Phonetics **11**, 373–382.

Lehiste, I. (**1970**). *Suprasegmentals* (MIT, Cambridge, MA), pp. 83 and 125.

McConnell-Ginet, S. (**1978**). "Intonation in a man's world," Signs **3**, 541–559.

McDonough, J. (**2002**). "The prosody of interrogative and focus constructions in Navajo," in *Formal Approaches to Functional Phenomena*, edited by A. Carnie and H. Harley (John Benjamins, Amsterdam), pp. 1–21.

Moore, C. B., and Jongman, A. (**1997**). "Speaker normalization in the perception of Mandarin Chinese tones," J. Acoust. Soc. Am. **102**, 1864–1877.

Mullennix, J. W., Johnson, K. A., Topcu-Durgun, M., and Farnsworth, L. M. (**1995**). "The perceptual representation of voice gender," J. Acoust. Soc. Am. **98**, 3080–3095.

Owren, M. J., Berkowitz, M., and Bachorowski, J.-A. (**2007**). "Listeners judge talker sex more efficiently from male than from female vowels," Percept. Psychophys. **69**, 930–941.

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

Schwartz, M. F. (**1968**). "Identification of speaker sex from isolated, voiceless fricatives," J. Acoust. Soc. Am. **43**, 1178–1179.

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. (**2007**). "Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled," J. Acoust. Soc. Am. **122**, 3628–3639.

van Bezooyen, R. (**1984**). *Characteristics and Recognizability of Vocal Expressions of Emotion* (Foris, Dordrecht), pp. 57–74.

van Dommelen, W. A., and Moxness, B. H. (**1995**). "Acoustic parameters in speaker height and weight identification: Sex-specific behaviour," Lang Speech **38**, 267–287.

Weinberg, B., and Bennett, S. (**1971**). "A study of talker sex recognition of esophageal voices," J. Speech Hear. Res. **14**, 391–395.

Whiteside, S. P. (**1998**). "Identification of a speaker's sex: A fricative study," Percept. Mot. Skills **86**, 587–591.

Wier, C. C., Jesteadt, W., and Green, D. M. (**1977**). "Frequency discrimination as a function of frequency and sensation level," J. Acoust. Soc. Am. **61**, 178–184.

Wong, P. C. M., and Diehl, R. L. (**2003**). "Perceptual normalization for inter- and intratalker variation in Cantonese level tones," J. Speech Lang. Hear. Res. **46**, 413–421.