JASA ARTICLE

Perception of gender in children's voices

Santiago Barreda[1,a)] and Peter F. Assmann[2]
[1]Department of Linguistics, University of California, Davis, California 95616, USA
[2]School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, Texas 75080, USA

ABSTRACT:
To investigate the perception of gender from children's voices, adult listeners were presented with /hVd/ syllables, in isolation and in sentence context, produced by children between 5 and 18 years. Half the listeners were informed of the age of the talker during trials, while the other half were not. Correct gender identifications increased with talker age; however, performance was above chance even for age groups where the cues most often associated with gender differentiation (i.e., average fundamental frequency and formant frequencies) were not consistently different between boys and girls. The results of acoustic models suggest that cues were used in an age-dependent manner, whether listeners were explicitly told the age of the talker or not. Overall, results are consistent with the hypothesis that talker age and gender are estimated jointly in the process of speech perception. Furthermore, results show that the gender of individual talkers can be identified accurately well before reliable anatomical differences arise in the vocal tracts of females and males. In general, results support the notion that the transmission of gender information from voice depends substantially on gender-dependent patterns of articulation, rather than following deterministically from anatomical differences between male and female talkers. © *2021 Acoustical Society of America*.
https://doi.org/10.1121/10.0006785
(Received 30 December 2020; revised 28 September 2021; accepted 30 September 2021; published online 23 November 2021)
[Editor: Jody Kreiman] Pages: 3949–3963

I. INTRODUCTION

Listeners can identify the gender of adult talkers with a high degree of accuracy. There is general agreement that this accuracy is driven by differences in the mean fundamental frequency (f_o, related to the length and mass of the vocal folds) and mean formant frequencies (related to the length of the vocal tract) produced by adult male and female talkers (Bachorowski and Owren, 1999; Hillenbrand and Clark, 2009). However, the anatomical variation that is largely responsible for the acoustic differences between adult men and women does not reliably emerge until after puberty (Story *et al.*, 2018; Vorperian *et al.*, 2009). Despite this, talker gender can be identified from the speech of prepubescent children at a greater-than-chance level. For example, Weinberg and Bennett (1971) report accuracy rates around 75% for spontaneous speech samples from 5 to 6 year old children, and Amir *et al.* (2012) found sex recognition rates averaging around 82% for isolated vowels and sentences produced by children between 8 and 18 years. Perry *et al.* (2001) presented adult listeners with CVC syllables produced by children aged 4, 8, 12, and 16 years and found perceptual differentiation at better than chance levels of male and female voices as young as four years of age.

The study described next is an investigation into the acoustic basis of gender perception from children's voices, with two primary goals. First, we wish to investigate the accuracy of gender identification from children's voices, and the way that this varies as a function of children's age and listening context. Second, we wish to explore the acoustic basis of listener's gender judgments, and especially the way that this may vary as a function of the age of the talker, and knowledge of the age of the talker.

A. Sex and gender in the voices of children

The perception of gender in children's voices presents an opportunity to untangle the anatomical and performative aspects of gender in the human voice. We may distinguish between sex, referring primarily to biological/anatomical characteristics, and gender, a set of practices that individuals engage in to maintain divisions between social categories such as "male" and "female" (Munson and Babel, 2019, p. 501). Zimman (2018) presents two general views of gender differences in voice: the essentialist view and the constructivist view. The essentialist view suggests that differences between male and female voices follow necessarily from anatomical differences between males and females. In this view "Biological sex is frequently the first-line explanation for any difference between women's and men's voices […] The concepts of sex and gender are often not distinguished, and typically, only two gender categories are recognized despite the fact that human sexual variation encompasses both bodies and identities that fall outside that binary" (Zimman, 2018, p. 3). In contrast, the constructivist view sees gender as a social practice, interpreting "differences between women, men, or any other gender group as socially learned rather than anatomically determined, at least in the absence of strong evidence to the contrary" (Zimman, 2018, p. 3).

a)Electronic mail: sbarreda@ucdavis.edu

The vocal tract morphology of boys and girls does not begin to differ reliably until puberty, between approximately 12–14 years of age (Vorperian *et al*., 2011; Fitch and Giedd, 1999). It follows that, from a strict essentialist perspective, talker gender should not be readily identifiable from speech until after puberty. Despite the absence of reliable anatomical differences between boys and girls before puberty, several studies have found consistent acoustic differences, since at least as far back as Sachs *et al*. (1973). For example, Lee *et al*. (1999) and Perry *et al*. (2001) reported significantly lower formant frequencies in boys compared to girls for vowels in /hVd/ syllables. Perry *et al*. (2001) showed that adult listeners distinguish boys from girls as young as four years of age from these recordings, and argued that such judgments are based on differences in average formant frequencies. Fitch and Geidd (1999) note, "We found no evidence for appreciable sex differences in children, suggesting that the clearly discriminable differences in girls' and boys' voices […] are primarily due to behavioral, not anatomical, differences" (p. 1515). In other words, the gender of young talkers can be identified when these talkers transmit information about their gender identity through their speech gestures, independently of the average physical differences in the vocal tracts of male and female talkers.

More recently, studies have shown that young children are sensitive to gendered voice characteristics, and can modify their speech to be perceived as more masculine or more feminine (Cartei *et al*., 2019a; Cartei *et al*., 2019b). When instructed to impersonate stereotypically male or female child characters, children can alter their voices to modulate the degree to which they are perceived as male or female, and they do this by raising or lowering their average fundamental frequency and formant pattern. These studies indicate that children form gender stereotypes well before the emergence of reliable differences in vocal tract anatomy and that they have the ability to change the gendered properties of their own speech to sound more masculine or feminine.

Cartei and Reby (2013) used a speech analysis/resynthesis method to modify the spectrum envelope of the speech from four 8-year old children (two males and two females). By systematically shifting the spectrum envelope along the frequency scale, they created several continua and asked adult listeners to label the voices as male or female or judge the masculinity/femininity of the voice using a rating scale. The spectrum envelope shifts employed in the stimulus design affect the spacing of formants in each vowel sound, which results in the impression of vocal-tract length variation across the continua. Results show that the probability that listeners report a male (or masculine) voice varied along each continuum in relation to the vocal-tract length implied by the acoustics of the stimulus token. Although the experiments did not explore the use of f_o, phoneme-specific variation, or the role of age in gender perception, the results establish that acoustic cues related to vocal-tract length are used by listeners when determining gender in the voices of children.

B. The role of age information in identifying the gender of children from voice

In a recent paper (Barreda and Assmann, 2018), we reported that adult listeners can provide reasonably accurate estimates of talker age (within ±1.8 years, on average) from isolated syllables spoken by children. In that experiment, one group of listeners was provided with information about talker gender while the other was not. Results suggested that both sets of listeners used acoustic cues in a gender-dependent manner, regardless of whether they were provided with information about talker gender. In cases where talker gender was not explicitly provided, listeners appeared to 'guess' the gender of the talker and adjust their use of acoustic cues. This behavior may be in response to the ambiguity arising from overlap in the speech acoustics of older female and younger male voices. This ambiguity can potentially be resolved by attending to age information when determining talker gender from speech. For example, the probability of observing a male talker with an f_o of 200 Hz is substantially higher for 11-year-old talkers relative to 18-year-old talkers.

C. Identification in syllables and sentences

Although talker gender can be accurately identified from isolated vowels, fricatives, and syllables produced by adults (Ingemann, 1968; Lass *et al*., 1976; Schwartz, 1968; Smith, 2016), gender information in adult voices is more robust in longer stretches of speech. For example, Hillenbrand and Clark (2009) found that scaling f_o and formant frequencies to simulate a talker of the opposite sex was more likely to induce changes in perceived gender for isolated syllables than for full sentences. Amir *et al*. (2012) report that the gender of children is also identified more accurately for sentences as opposed to syllables, although their results suggested that this effect may be more pronounced for male children.

The improved performance for sentences over syllables may simply be because sentences provide a better opportunity to estimate the acoustic parameters relevant for gender perception. Alternatively, sentences may better convey prosodic information (e.g., speech rhythm, f_o contour) that can potentially be useful to listeners in gender identification (Clopper and Smiljanic, 2011). Finally, there is the possibility that sentences could help gender identification from the speech of children indirectly, by suggesting a younger or older talker. For example, spectral and temporal variability in production (Gerosa *et al*., 2007; Lee *et al*., 1999), and coarticulatory variability (Khwaileh, 2011) can decrease as a function of talker age. Since producing a sentence involves substantially more gestural planning and coordination than a single syllable, sentences may better convey talker maturity (and therefore age) to listeners.

II. METHODS

A. Listeners

Forty undergraduate students (31 females, 9 males) at the University of Texas at Dallas participated in the

https://doi.org/10.1121/10.0006785

experiment. All were native talkers of American English with normal hearing as determined by pure-tone screening at 500, 1000, 2000, and 4000 Hz. They were compensated with experimental research credits for their participation and provided written informed consent prior to the listening test.

B. Stimuli

The stimuli were taken from a children's speech database (Assmann *et al.*, 2008). For the syllable context, 140 talkers (five boys and five girls at each age level between 5 and 18 years) each contributed three syllables: /hid/ ("heed"), /hɑd/ ("hod") and /hud/ ("who would"). For the sentence context, the number of stimuli was reduced to preserve the overall time allotted to complete the experiment. A subset of three of the five talkers produced the three syllables in a carrier sentence ("Please say the word _____ again."). As a result, listeners in the syllable context heard 420 stimuli (5 talkers × 14 ages × 2 sexes × 3 vowels); those in the sentence context heard 252 stimuli (3 talkers × 14 ages × 2 sexes × 3 vowels).

C. Procedure

Half of the listeners (20) were assigned to listen to stimuli composed of isolated syllables; the other half listened to stimuli composed of sentences. Within each listening context, half of the participants (ten) were randomly assigned to receive information about the age of the talker on the computer screen following the presentation of each stimulus; the other half did not. This resulted in 2×2 between-subjects design for listening context (sentence vs syllable), and age-information (with and without).

Stimuli were presented monaurally in a double-walled sound booth at a mean level of 68 dB sound pressure level (SPL) (A-weighting), using the Tucker-Davis System 3 and RP2.1 hardware with Sennheiser HD-598 headphones. Stimuli were randomized along all stimulus dimensions and presented in a different random order to each listener. Following each stimulus, participants identified the talker as either "male" or "female" using response buttons, then assigned a confidence rating on a 5-point scale. The confidence ratings will not be discussed here. Prior to the main experiment listeners completed the hearing screen and a brief questionnaire, followed by 24 familiarization trials, using stimuli similar to those in the experiment but spoken by different talkers. Feedback was provided for the practice items but not for experimental trials. The experiment was self-paced and took about 50 min, with an optional break at midpoint.

D. Analysis

Two analyses were carried out, one investigating sensitivity and bias in gender judgments, and another investigating the use of acoustic information in the determination of talker gender. Both analyses were carried out using Bayesian multilevel-models fit in R (R Core Team, 2019) using the "brms" package (Bürkner, 2018).

1. Analysis of sensitivity and bias

The probability of observing a male response on a given trial was modeled as a function of the talker's self-identified gender (G, "male" or "female"), whether talker-age information was provided (I, with or without), and listening context (C, syllable or sentence), with all possible interactions included. The talker's true age was included as a "random" effect, and random by-age slopes were also calculated for the full set of "fixed-effect" predictors and interactions. Listener and talker were included as random effects, and a random by-listener effect for gender was also included. The formula used for the model is presented in Eq. (1),

$$P(response = \text{'male'}) \sim G*I*C + (G*I*C \,|\, Age) + (G \,|\, Listener) + (1 \,|\, Talker). \quad (1)$$

The random by-age slopes for G, I, and C (and their interactions) in Eq. (1) allow the model to represent age-related variation in sensitivity and bias. Age was included as a random effect in order to estimate the large number of involved parameters ($2 \times 2 \times 2 \times 14 = 112$ for the G:I:C:Age interaction), while avoiding negative outcomes that might arise from treating these as "fixed" effects. Although the term "random effects" has several inconsistent definitions, it is often used to refer to the estimation of parameters using "partial pooling," while "fixed" effects are estimated entirely independently (Gelman, 2005). Partial-pooling estimates take into account the values of the *other* levels in a given factor, making the appearance of spurious values less likely, and generally protecting from many of the negative outcomes related to the estimation of large numbers of parameters in more traditional models (Gelman *et al.*, 2012).

By modeling responses of "male" and including actual gender as a predictor, the analysis in Eq. (1) allows us to investigate sensitivity (the ability to separate categories) and bias (the tendency to respond to one category more than another) across the different listening conditions. As noted in DeCarlo (1998), the use of logistic regression to fit a signal-detection theory model yields equivalent results to the more traditional probit model (yielding a d' analysis). The logistic approach is preferred here because of the easier interpretability of coefficients when these are expressed in log-odds. In the parametrization presented in Eq. (1), the gender main effect (and any term interacting with gender) reflects sensitivity, and any coefficients that *do not* interact with gender reflect bias (DeCarlo, 1998).

2. Acoustic analysis of sex judgments

The acoustic model is based on information collected from the vowels contained by the stimuli in the syllables condition. Formant frequencies were measured every 2 ms using an automatic formant tracking program (Nearey *et al.*, 2002), and fundamental frequency was estimated with 1-ms resolution (Kawahara *et al.*, 2005). In each case, syllables

were represented by median values across 20 frames centered at the vowel midpoint. All formant and f_o tracks were visually verified for correctness. Around 10% of tokens (37/420) required manual correction of one or more formants, which was carried out using TrackDraw (Assmann *et al.*, 1994).

Modeling acoustic information from whole sentences results in several complications including the representation of dynamic information, the aggregation of evidence across segments, and the updating of initial inferences. To our knowledge, none of these issues have widely accepted solutions. Therefore, as a starting point, we assume that roughly the same acoustic information is used in largely the same manner for syllables and sentences (i.e., that estimates are largely consistent across conditions). We expect that the acoustic analysis for the syllable context will generally apply to the sentence context, with the caveat that in practice dynamic acoustic information likely also plays an important part in gender perception (discussed further in Sec. IV E).

We are interested in three general questions: (1) What acoustic information determines apparent talker gender in the voices of children? (2) Does the use of acoustics vary in an age-dependent manner? (3) Is this affected by providing information about the age of the talker? To address these questions, we fit the model in Eq. (2), which predicts the probability of observing a response of "male" as a function of the acoustic predictors in Table I (represented by the vector X) and vowel category (V), both interacting with talker-age information (I). All acoustic predictors were first standardized across all talkers so that each predictor had a mean of 0 and a standard deviation of 1 (original standard deviations are provided in Table I).

$$P(response = \text{`}male\text{'}) \sim (X + V) * I + (X * I \,|\, Age) + (X + V \,|\, Listener) + (1 \,|\, Talker). \quad (2)$$

The model includes random by-talker intercepts, and random by-listener slopes for each acoustic predictor (and vowel category). The model treats age as a random effect, includes random by-age slopes for acoustics, and allows these to vary by age-information condition. As with the model outlined in Sec. II D 1, including age as a random effect results in protection from spurious effects through the application of partial-pooling to parameter estimates.

The measures presented in Table I were chosen based on evidence in the literature indicating their capacity to affect listeners' judgments of talker gender from speech. Source-related predictors were estimated using VoiceSauce (Shue *et al.*, 2009), formant frequencies (F1, F2, and F3) were estimated using the Nearey formant tracker (Nearey *et al.*, 2002) and fundamental frequency (f_o) was estimated using STRAIGHT (Kawahara *et al.*, 2005). All acoustic parameters were sampled at the midpoint of the vocalic portion of the syllable. Figure 1 presents averages for male and female talkers in each age group for selected acoustic predictors.

Between-talker variation in average formant frequencies (i.e., vocal-tract length) was operationally defined using the logarithm of the geometric-mean formant frequency produced by a talker across all their vowel tokens ($\overline{G}_{talker}$, Nearey, 1978; Nearey and Assmann, 2007). As noted in Barreda (2020), when talkers are sampled from a single dialect, differences in their $\overline{G}_{talker}$ will primarily reflect differences in their vocal tract length.[1] Vowel category was included as a (sum-coded) factor to express information about predictable variation in acoustic properties between vowels, including the average F1, F2, and F3 values of different phonemes. To minimize correlations in the data, individual formant frequencies (F1, F2, F3) were log-transformed and then centered within vowel-category (Hillenbrand and Clark, 2009). These values will reflect the position of each token relative to the mean of the phoneme-specific distribution seen in Fig. 1(a) (within-phoneme variation).

Using log-transformed Hertz (log-Hz) as a unit of measurement means that effects will be proportional to the baseline value of the formant. For example, an increase in 0.1 log Hz indicates an increase in Hertz values by a factor of 1.105 [exp (0.1)], approximately 10%. In contrast, an increase in 100 Hz can suggest substantially different proportional changes based on whether the underlying formant began at 300 Hz (33% increase) or 1100 Hz (9% increase). Thus, the use of log-Hz allows the model to code proportional changes in formant frequencies (and derived

TABLE I. Acoustic predictors used in modeling sd of each acoustic predictor across all data.

Abbr.	Description (source)	sd (unit)
f_o	natural logarithm of the fundamental frequency (Hillenbrand and Clark, 2009)	0.29 (log-Hz)
CF1	category-centered log-F1 (Hillenbrand and Clark, 2009)	0.11 (log-Hz)
CF2	category-centered log-F2 (Hillenbrand and Clark, 2009)	0.16 (log-Hz)
CF3	category-centered log-F3 (Hillenbrand and Clark, 2009)	0.15 (log-Hz)
$\overline{G}_{talker}$	mean of the logarithms of F1, F2 and F3 across all vowels (Assmann *et al.*, 2008)	0.13 (log-Hz)
CPP	Cepstral pitch prominence (Hillenbrand *et al.*, 1994)	4.0 (decibels)
HNR	Harmonics-to-noise ratio (de Krom, 1993)	13.0 (decibels)
H1H2c	Corrected magnitude-difference between harmonics 1 and 2 (Iseli *et al.*, 2007)	7.5 (decibels)
H1A1c	Corrected magnitude-difference between harmonic 1 and F1 peak (Iseli *et al.*, 2007)	8.3 (decibels)
H1A3c	Corrected magnitude-difference between harmonic 1 and F3 peak (Iseli *et al.*, 2007)	4.0 (decibels)

https://doi.org/10.1121/10.0006785

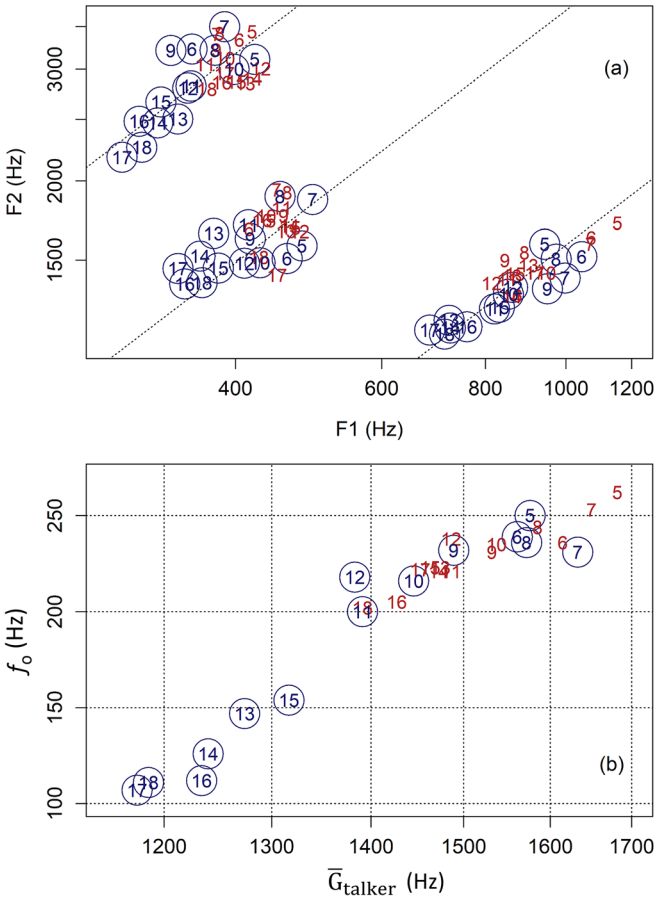

FIG. 1. (Color online) (a) Average F1 and F2 values for each vowel (upper left, /i/; lower left, /u/; lower right, /ɑ/) for each age and gender group. (b) Mean $\overline{G}_{talker}$ and f_o for each age and gender group. Numbers indicate age groups, males in blue and females in red. Male values are presented in circles.

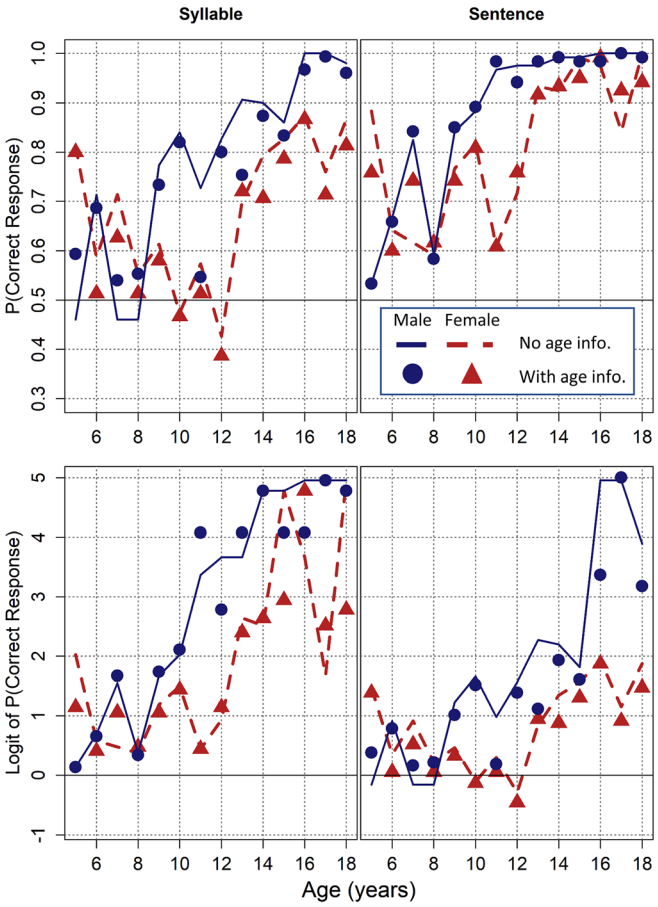

FIG. 2. (Color online) Probability of observing a correct gender identification, pooled across all listeners, presented by age and gender group. Each column presents the same information (left, syllable context; right, sentence context): the top row presents probabilities while the bottom row presents the logit of the same values.

measures) of the kind associated with differences in vocal-tract length.

III. RESULTS

A. Analysis of sensitivity and response bias

Gender was identified more accurately from sentences than syllables (84% vs 72% correct), but information about talker age actually had a small *negative* effect on average accuracy (76% with age information, 78% without age information). Gender was identified more accurately for older than younger talkers, particularly for teenaged compared to prepubescent talkers (Fig. 2). In addition, listeners were generally more correct in identifying the gender of male talkers than that of female talkers (82% vs 73% correct), reflecting a possible bias towards identifying talkers as male.

Responses were analyzed using the model described in Sec. II D 1. In this model, coefficients interacting with gender reflect increasing sensitivity (the ability to distinguish categories), with positive values indicating a greater separability of the categories. Coefficients *not* interacting with gender represent bias (the overall tendency to select one category over another), with positive coefficients reflecting an increasing tendency to respond "male." In our logistic regression model, sensitivity measures reflect the difference between hit and false alarm rates in log-odds, so that a value of d represents an expected hit-rate of $1/(1+e^{-d/2})$ and an expected false alarm rate of $1/(1+e^{d/2})$, in the absence of any bias. An intercept of c reflects a shift of c log-odds in both hits and false alarms relative to the zero-bias case, reflecting a c log-odds increase in the overall probability of observing a response of "male" [e.g., hit rate $= 1/(1+e^{-d/2+c})$].

We will distinguish fixed predictors and age-varying predictors (the age-related "random" effects). Fixed predictors [Fig. 3(a)] reflect overall differences in sensitivity and bias across listening conditions without considering age-related variation in these parameters. Listening context and age-information were coded with treatment coding, with isolated syllables and no age-information serving as reference levels (i.e., the intercept represents syllables with no age-information). Results indicate that listeners may have a very small bias towards responding male overall [Intercept $= 0.51$, standard deviation (sd) $= 0.28$, 95% highest density

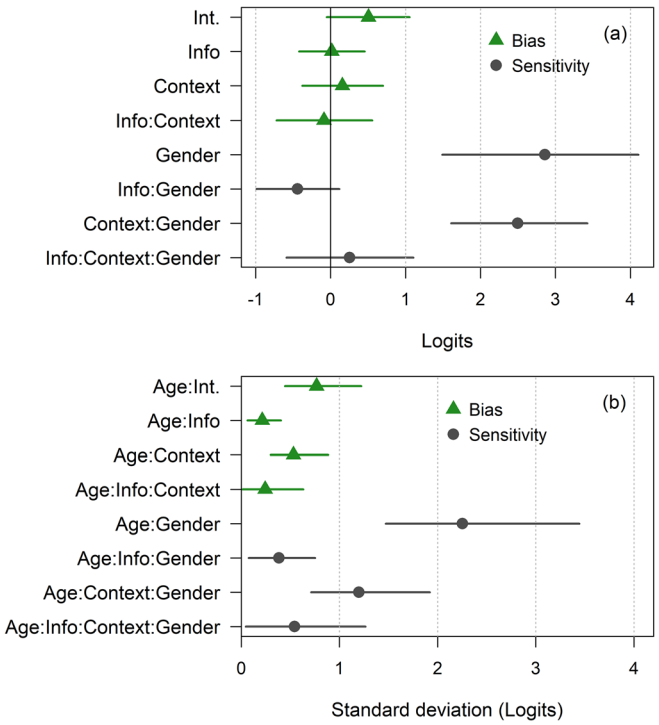

FIG. 3. (Color online) (a) Posterior distributions of estimates of the fixed (marginal) effects of different predictors. (b) Posterior distributions of estimates of variance components related to age-related variation. Points indicate means, bars indicate 95% highest-density intervals.

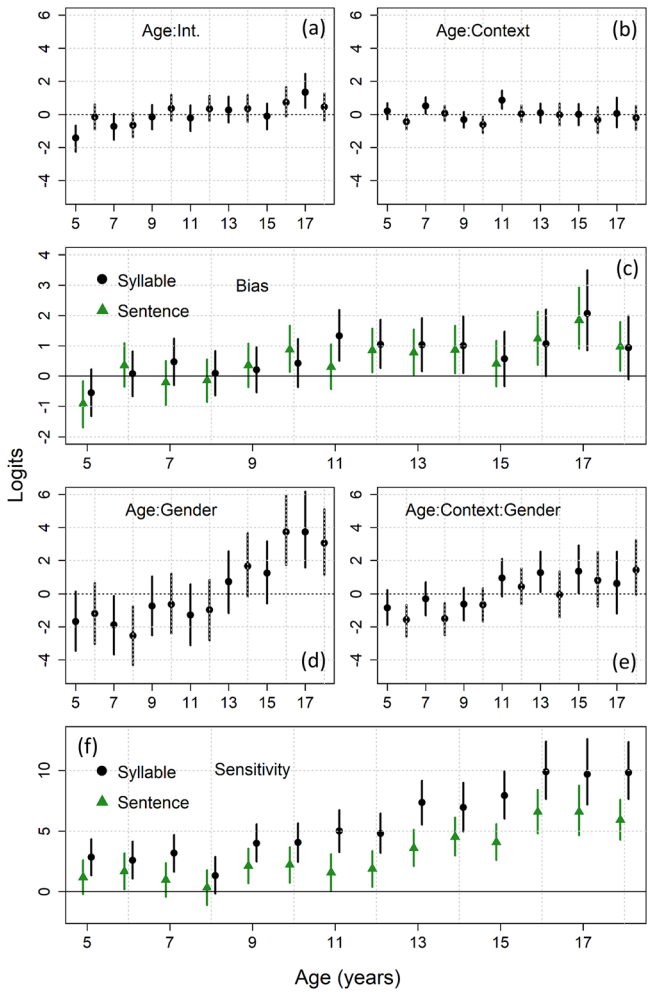

FIG. 4. (Color online) Distribution of selected age-varying predictors (i.e., by-age random effects). Points indicate means, bars indicate 95% highest-density intervals. Effects in (c) and (f) represent the sum of the appropriate fixed predictors [from Fig. 3(a)] to the age-varying terms presented in the row immediately above.

interval (HDI) = (−0.05,1.05)], however, there is no apparent variation in response bias according to age-information or listening context. Although listeners were quite good at distinguishing male and female talkers in syllables [Gender = 2.86, sd = 0.66, 95% HDI = (1.49, 4.11)], sensitivity was nearly twice as large in the sentence context [Context:Gender = 2.50, sd = 0.46, 95% HDI = (1.61, 3.43)]. Age information did not improve sensitivity, and may have actually had a small negative effect [Info:gender = –0.44, sd = 0.28, 95% HDI = (–0.99, 0.12)].

To investigate the age-varying parameters, we considered the standard deviation of different groups of predictors (Gelman, 2005), presented in Fig. 3(b). Groups of coefficients with large standard deviations vary substantially from each other and thus have a larger effect on observed outcomes. The age:intercept (i.e., the by-age random intercepts) and age:context terms reflect variation in response bias according to talker age, and variation in this according to listening context (syllables vs sentences). There does not appear to be any age-related effect for talker-age information. Similarly, the large values for age:gender and age:context:gender indicate substantial variation in sensitivity by age, and these also vary by listening context. Since all of the coefficient groups involving age information have small standard deviations, with many values near zero, explicitly providing age information does not seem to have had a notable effect on age-dependent listener behavior. Figure 4 presents the groups of age-varying coefficients found to have non-trivial amounts of variation. Figures 5(c) and 5(f) present comparisons of predicted bias and sensitivity at different ages, across the two listening contexts.

The biases in Fig. 4(c) may at first appear to exhibit a positive linear trend. However, apart from the negative values at five years, the bias parameters vary around zero until about 11 years of age, after which time they are mostly positive. This suggests a growing tendency to respond "male" for post pubescent voices (discussed further in Sec. IV D). A similar pattern is evident in the listener sensitivities, which are reasonably stable until they begin to increase after ages 11–12. Although there is a clear benefit to the sentence listening context and much higher sensitivity after puberty, listeners were still able to identify talker gender at a higher than chance level for children 5–8 years old based only on isolated syllables. For example, the predicted average sensitivity for ages 5–8 in the syllable context was $d = 1.04$ [sd = 0.41, 95% HDI = (0.25, 1.84)], which indicates an expected accuracy of 63% [95% HDI = (53%, 71%)] even for these very young children.

https://doi.org/10.1121/10.0006785

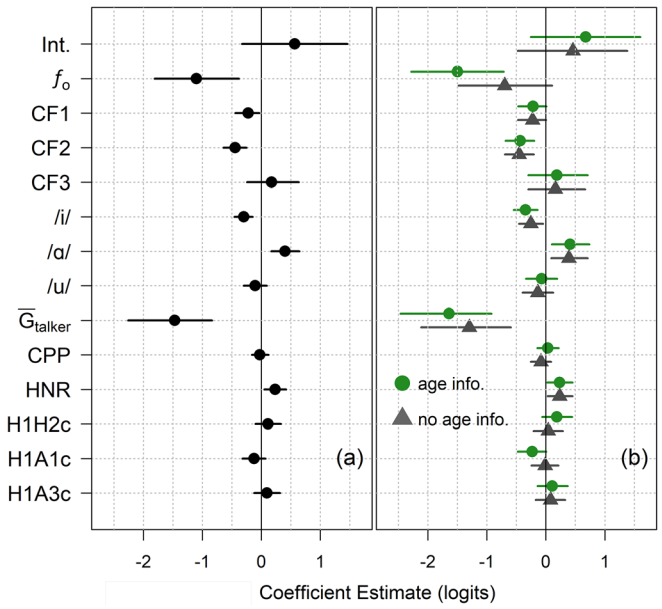

FIG. 5. (Color online) (a) Values of fixed acoustic parameters. (b) Effects are contrasted across the age-information conditions. Points indicate means, bars indicate 95% highest-density intervals. Posterior means and credible intervals for all parameters are available in supplemental material.

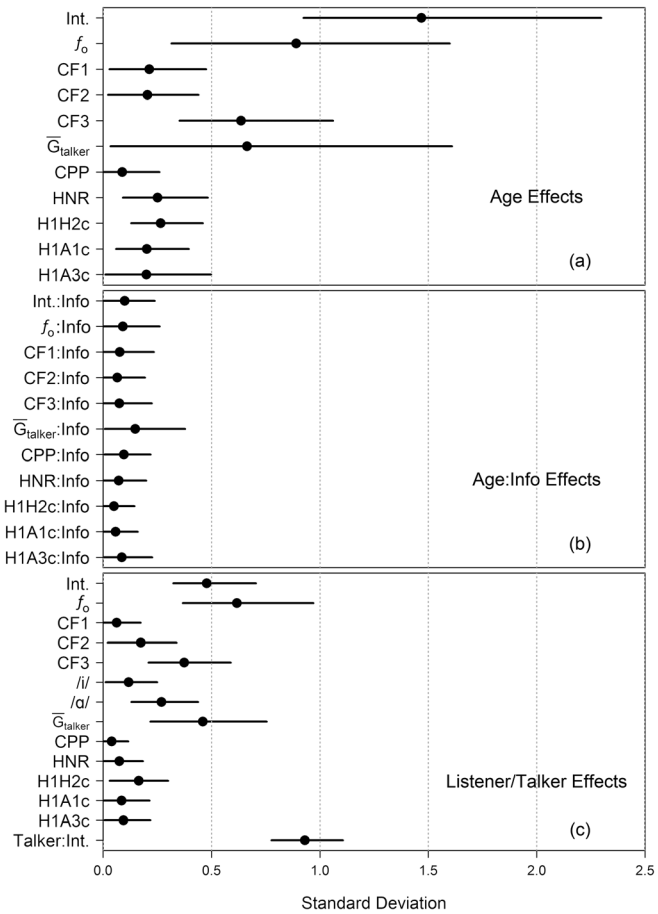

FIG. 6. Posterior distributions of estimates of the standard deviation of different variance components. Points indicate means, bars indicate 95% highest-density intervals.

B. Acoustic analysis

Figure 5(a) presents the fixed parameter estimates for each acoustic predictor, reflecting average use of each cue across all ages. Results indicate that the strongest spectral effect by far was for $\overline{G}_{talker}$ with weaker, but consistent, effects for centered F1 and F2 (CF1, CF2), and for vowel category. The predictors representing individual formants specifically capture the marginal effects of these *when all other formants are held constant*. Thus, the $\overline{G}_{talker}$ predictor represents the effect of coordinated shifts to all formants (i.e., vocal-tract length differences) that cannot be explained by added effects of the individual formant predictors.

There was a strong effect for f_o, however the remaining source predictors had very weak effects on responses, and all but HNR are difficult to distinguish from zero. In Fig. 5(b), we see that age-information had very small effects on the acoustic predictors for the most part, with a small increase in the magnitude of $\overline{G}_{talker}$ [Mean = 0.34, sd = 0.30, 95% HDI = (–0.275, 0.936)] and a moderate increase in the values of f_o [Mean = 0.80, sd = 0.34 95% HDI = (0.084, 1.48)].

1. Age-dependent use of acoustic information

To investigate the age-dependent use of acoustics, we considered the standard deviation of each bundle of random-effects terms (as in Sec. III A). A large standard deviation for a coefficient, for example the f_o term in Fig. 6(a), indicates substantial variation in the value of a parameter across ages. As seen in Fig. 6(a), there is substantial age-related variation in the intercept, f_o, CF3, and $\overline{G}_{talker}$, and a very small amount of variation in HNR, H1H2c, and H1A1c. Figure 6(b) presents the variation in the age-information:acoustics interactions, which are all near zero. This suggests there are no meaningful differences in age-dependent use of acoustic information across the age-information conditions. In other words, listeners do use some acoustic cues in an age-dependent manner, but having age information explicitly provided to them made no difference to this behavior. As a result, the age-dependent use of acoustics will be considered jointly across both age-information conditions.

Figure 7 presents the distribution of the age-varying acoustic predictors that showed substantial variation in Fig. 6(a). Intercept values decrease as a function of talker age, likely reflecting the fact that a talker with average f_o and $\overline{G}_{talker}$ for this dataset would have relatively high frequencies for an adult male, but low frequencies for a young male. Thus, if we fix the acoustic properties of a voice at the average, the probability that the talker is male will decrease as a function of talker age. Although this behavior would account for the pattern of responses seen in Fig. 7, it assumes that age is known, and listeners provided responses consistent with this age dependent behavior even in the absence of explicit information about talker age.

The coefficients for f_o and $\overline{G}_{talker}$ share a similar pattern, with small negative values increasing in magnitude at approximately 12–13 years of age. Although substantially noisier (and smaller in magnitude), the age-related patterns for HNR and H1H2c are suggestive of the same general

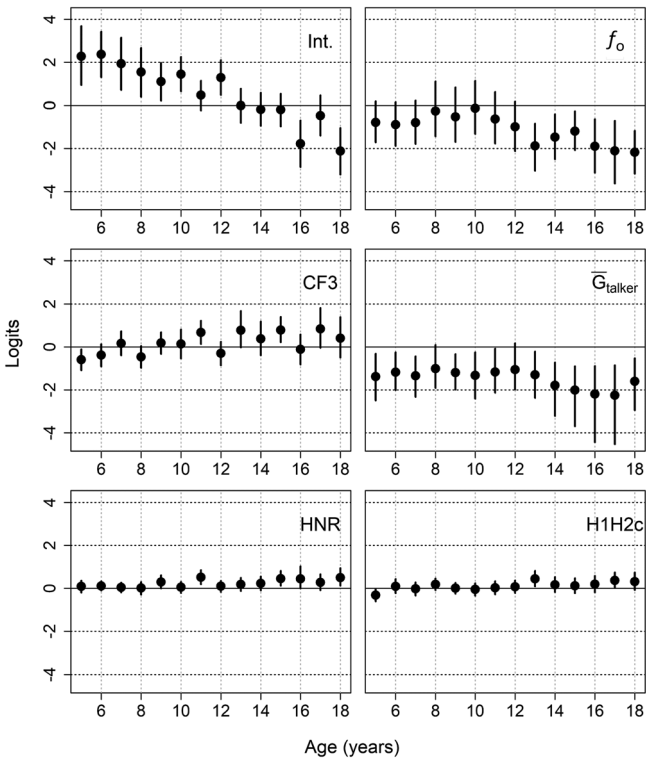

FIG. 7. Effects for each acoustic predictor as a function of age. Points indicate means, bars indicate 95% highest-density intervals. Age-dependent effects are found by adding the marginal effect for each predictor (presented in Fig. 5) to the age-dependent effects for that predictor (e.g., f_o + Age: f_o to find age-dependent values for f_o).

trend, with coefficient magnitudes increasing at approximately 12–13 years of age. The age-related variation in CF3 features no clear pattern, with changes in sign across successive years. Since its marginal, fixed parameter estimate is also quite noisy and close to zero, the age-dependent variation in CF3 may reflect a combination of noise and variation in the characteristics of the talkers in our age groups. The effect for H1A1c is not presented as its small deviations from zero seem to reflect only noise, with no consistent trend.

IV. DISCUSSION

Our results show that listeners can distinguish males from females at levels better than chance even for young children, and that identification accuracy is higher for sentences compared to isolated syllables. Providing listeners with information about talker age did not have a strong effect on listener sensitivity or bias; however, all listeners appeared to use talker age information in their judgments of talker gender. In this section, we will discuss the implications of these findings for the estimation of vocal-tract length from speech acoustics, the joint estimation of speech categories and talker indexical characteristics, and the transmission of gender information from speech.

A. Vocal-tract length estimation in gender perception

As seen in Fig. 5 and Table II, within-category variation in the formant frequencies (CF1, CF2) and phoneme-specific variation between vowel categories can both have weak, but consistent, effects on the perception of talker gender. In both cases, lower formants are associated with the perception of male talkers, conforming to previous reports that vowels with lower formants tend to be associated with larger talkers (Barreda, 2017). However, although the effects for within-speaker variation in formant patterns are consistent, the effect is much smaller than it might be. For example, in our data /i/ has average F1, F2, and F3 frequencies that are 86%, 180%, and 121% of the value of the average F1, F2, and F3 frequencies of /u/. Despite these large differences in their formant patterns, there is only a 0.51 logit [sd = 0.21, 95% HDI = (0.11, 0.92)] difference in the probability of a "male" identification between /i/ and /u/. Compare this to the expected effect of a 14% change in $\overline{G}_{talker}$, which is more than three times larger (–1.47 logits). Similarly, despite standard deviations of roughly equal magnitude to that of $\overline{G}_{talker}$ (see Table I), the effects for CF1 (−0.22) and CF2 (−0.44) are substantially smaller than that of $\overline{G}_{talker}$.

TABLE II. Posterior means for coefficients from the acoustic model [averaged across age-information conditions, Fig. 5(a)]. The sd of posterior distributions relate to the standard error of these estimates, and the values indicated by 2.5% and 97.5% reflect the bounds of the 95% highest-density credible interval for the parameter.

Parameter	Mean	sd	2.5%	97.5%
Intercept	0.57	0.45	−0.33	1.46
f_o	−1.1	0.36	−1.81	−0.38
CF1	−0.22	0.1	−0.44	−0.03
CF2	−0.44	0.1	−0.64	−0.25
FC3	0.17	0.22	−0.24	0.63
/i/	−0.3	0.08	−0.46	−0.15
/ɑ/	0.4	0.12	0.17	0.64
/u/	−0.1	0.1	−0.3	0.09
$\overline{G}_{talker}$	−1.47	0.36	−2.26	−0.84
CPP	−0.02	0.07	−0.16	0.12
HNR	0.23	0.09	0.05	0.42
H1H2c	0.11	0.11	−0.1	0.33
H1A1c	−0.12	0.1	−0.32	0.07
H1A3c	0.09	0.11	−0.12	0.32

Overall, our results show that listeners are more influenced by the average spectral characteristics of a talker (e.g., $\overline{G}_{talker}$) than they are by independent variation in the absolute formant frequencies of a given token. This leads to an apparent paradox: the best predictor of talker-gender judgments ($\overline{G}_{talker}$) is not directly present in the signal. A plausible interpretation for this is that listeners make use of the absolute formant pattern to estimate a latent talker-dependent variable (such as $\overline{G}_{talker}$), which is then used to infer talker gender and potentially other talker indexical characteristics (Barreda, 2020; Nearey and Assmann, 2007; Turner *et al.*, 2009). Here, we outline some ways that listeners might recover this information from individual tokens.

To estimate talker characteristics in a stable manner despite between-phoneme variation, listeners require a

https://doi.org/10.1121/10.0006785

talker-dependent, phoneme-independent value (e.g., $\overline{G}_{talker}$). However, listeners only have access to the phoneme-dependent spectral information in a given token (e.g., the mean of log-F1, log-F2, and log-F3 for a token ($\overline{G}_{token}$). Using token characteristics directly as an estimate of $\overline{G}_{talker}$ (or a similar statistic) can yield reasonable estimates, with errors on the order of approximately 10% of values (Johnson, 2020; Lammert and Narayanan, 2015). Estimates of $\overline{G}_{talker}$ based on single tokens that do not control for phonetic content may include substantial phoneme-dependent biases in their estimates of $\overline{G}_{talker}$ (Barreda and Nearey, 2018). As an example of this approach, we use $\overline{G}_{talker}$ directly to estimate $\overline{G}_{talker}$ [Fig. 8(a)], resulting in a mean absolute prediction error of 10% and estimates that have a correlation of 0.7 with $\overline{G}_{talker}$ [Fig. 8(a)]. These estimates of $\overline{G}_{talker}$ feature large biases (seen as intercept shifts) that are entirely predictable on the basis of the vowel category used to estimate the parameter.

Barreda (2017) outlines a simple "pattern correction" method to estimate $\overline{G}_{talker}$ more accurately in cases where the vowel category is known.[2] Dialects will have systematic, phoneme dependent variation in $\overline{G}_{talker}$ based on the phoneme uttered [as seen in Fig. 8(a)]. For example, the average $\overline{G}_{talker}$ across all talkers in this experiment was 1558, 1546, and 1260 Hz for /i/, /ɑ/, and /u/, respectively [where frequency in Hz = exp ($\overline{G}_{talker}$)]. The average $\overline{G}_{talker}$ across all vowels was 1447 Hz, meaning that the vowel-specific formant averages are, on average, 108%, 107%, and 87% as large as $\overline{G}_{talker}$ when expressed in Hz. In fact, we can see these tendencies reflected in the between-phoneme intercept shifts in Fig. 8(a). Thus, if we simply multiply any given $\overline{G}_{talker}$ by the correct phoneme-specific scale adjustment, we may recover reasonable estimates of $\overline{G}_{talker}$. This process is analogous to adjusting the intercepts of the lines in Fig. 8(a) based on the expected vowel-specific distance to the overall mean intercept. This adjustment results in a mean absolute prediction error of 4% [Fig. 8(b)], and produces estimates that have a correlation of 0.90 with $\overline{G}_{talker}$. Thus, we see that even this very simple adjustment can substantially reduce prediction error when estimating average talker characteristics (e.g., $\overline{G}_{talker}$) from a single token.

More complicated pattern-correction methods exist that can accurately estimate $\overline{G}_{talker}$ from single tokens even when the vowel category is not known. For example, Nearey and Assmann (2007) present a model (Method 6, p. 258) that uses knowledge of dialectal phoneme patterns to estimate a different "pattern-corrected" $\overline{G}_{talker}$ for each possible vowel-category (rather than only for the "known" category). These estimates are then considered together with information about $\overline{G}_{talker}$ ranges across the human population, and information about the covariance of $\overline{G}_{talker}$ and g0, in order to jointly estimate $\overline{G}_{talker}$ and vowel-category for each token. We created a model as described in Nearey and Assmann using acoustic measurements of our stimuli, and used this to classify vowel category and estimate $\overline{G}_{talker}$ for our vowel stimuli. This model was given information about stimulus F1, F2, F3, and f_o, but no direct information about average talker information (e.g., $\overline{G}_{talker}$), or about the true vowel category. This model was able to classify 99% (416/420) of tokens correctly, estimated $\overline{G}_{talker}$ to within a mean absolute error of 2.8%, and produced estimates with a correlation of 0.94 with the true $\overline{G}_{talker}$.

Barreda (2020) provides a conceptual framework to consider these "pattern corrections" in perception by comparing $\overline{G}_{talker}$ estimation and vowel recognition to size/shape estimation in the visual domain. In vision, distance scales the apparent size of objects up/down uniformly along all dimensions. As a result, the size of the retinal image of an object is a reliable cue to the distance of the object from the observer. Objects of different sizes can appear to be the same size when presented at different distances. As a result, when an observer estimates the distance of an object, they must consider both the apparent size (i.e., the retinal image) and the known (or estimated) true size of the object (Holway and Boring, 1941). This behavior is analogous to the joint estimation of phonetic content and vocal-tract

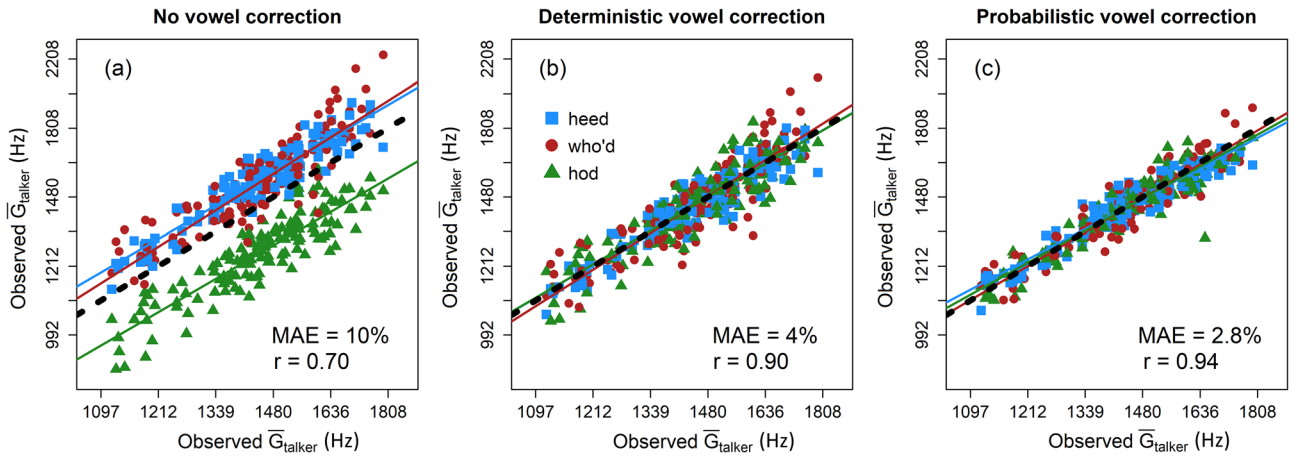

FIG. 8. (Color online) Distribution of predicted and observed $\overline{G}_{talker}$ for different estimation methods. Each solid line indicates the relationship between predicted and observed $\overline{G}_{talker}$ for a different vowel. Dotted lines indicate perfect prediction. (a) Estimation of $\overline{G}_{talker}$ directly from the formants with no vowel correction. (b) Correcting for the known vowel category results in a substantial improvement in estimation. (c) With more sophisticated $\overline{G}_{talker}$ estimation methods (as described in the text), extremely accurate estimates can be achieved from even a single token.

length. The vocal tract scales the formant pattern associated with a vowel phoneme up or down, so that measures of average formant frequency (e.g., $\overline{G}_{talker}$) are reliable cues to the size of the talker. However, since different phonemes (i.e., linguistic "objects") have different inherent "sizes," the listener must consider the properties of the object when estimating the VTL-related scaling from the spectral pattern. Essentially, an F1 of 500 Hz means substantially different things if it is observed for /i/, as opposed to for /ɑ/, and listeners appear to behave in a manner consistent with this knowledge.

B. Age-dependent use of acoustics

In Barreda and Assmann (2018) we suggested that listeners determine the talker's gender in order to help improve age perception from speech. Conversely, in this study, we find evidence of the reliance on talker-age information in gender perception. To evaluate the importance of the age-related variation in acoustics (Fig. 6), we investigated the predictive accuracy of variants of the model in Eq. (2), presented in Fig. 9. These variants contain the same fixed-effects structure, but differ in the random-effects terms excluded or included when making predictions. Rather than looking for the highest accuracy, we are interested in the most accurate prediction of listener responses. The full model [Fig. 9(a)] is able to predict perceived talker-gender with good accuracy; however, this includes both talker- and listener-specific adjustments to the model. Figure 9(c) indicates that a model including only age-related adjustments is still able to predict talker gender with good accuracy, and is still a reasonable reflection of listener behaviors. In contrast, a model that uses only the marginal effects for acoustics [Fig. 9(d)] without considering talker age offers very inaccurate prediction of listener judgements. Although the talker and listener adjustments improve the situation somewhat [Figs. 9(e) and 9(f)], the predictions are still substantially different from the observed listener judgments.

Figure 9 suggests that the age-dependent use of acoustics may be an important aspect of the accurate gender perception exhibited by listeners. As noted by Barreda and Assmann (2018), this is likely a result of the substantial overlap between older female and younger male talkers. Figure 10 presents the location of each talker in this experiment according to their average f_o and $\overline{G}_{talker}$, presented by age. As seen in Fig. 10, using a single boundary for all talkers leads to systematic errors as a function of talker age [see Fig. 9(d)], with a tendency to classify all younger talkers as female and all older talkers as male. In contrast, the shifting boundaries mode possible by the age-dependent models have a much greater chance of yielding correct gender identifications, and better reflect the responses provided by human listeners.

C. Joint estimation of age and gender

Recently, Barreda (2020) suggested that the perception of vowel quality may be inherently related to the perception of talker size (and other indexical characteristics) *via* their shared dependence on the apparent spectral-scaling associated with a vowel (indexed here using $\overline{G}_{talker}$). Next, we provide a sketch of how the perception of talker age and gender might naturally be related, though we do not make any claims about any specific implementation.

It has long been noted that the perception of vowel quality involves the estimation of a talker-dependent spectral scaling parameter ($\overline{G}_{talker}$, or an analogous measure) related to vocal-tract length (Nearey, 1978; Turner *et al.*, 2009), even if only as a by-product of speech perception. For example, consider the distribution of average productions for each talker group presented in Fig. 1(a), presented again in Fig. 11(a). In order to recognize a vowel, a listener

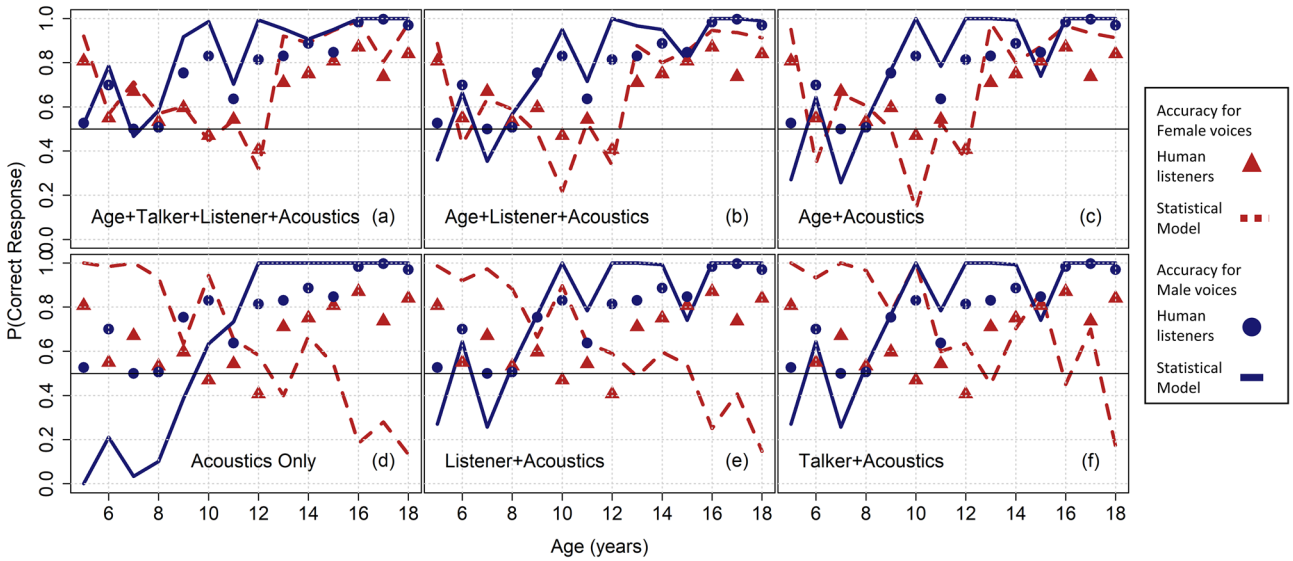

FIG. 9. (Color online) Accuracy of predictions made using models with the same fixed-effects structure, but differing in the included random-effects clusters (indicated in each plot). Prediction of male and female voices is presented separately. Model accuracy is summarized using the category implied by the mean posterior predicted log-odds for each token.

https://doi.org/10.1121/10.0006785

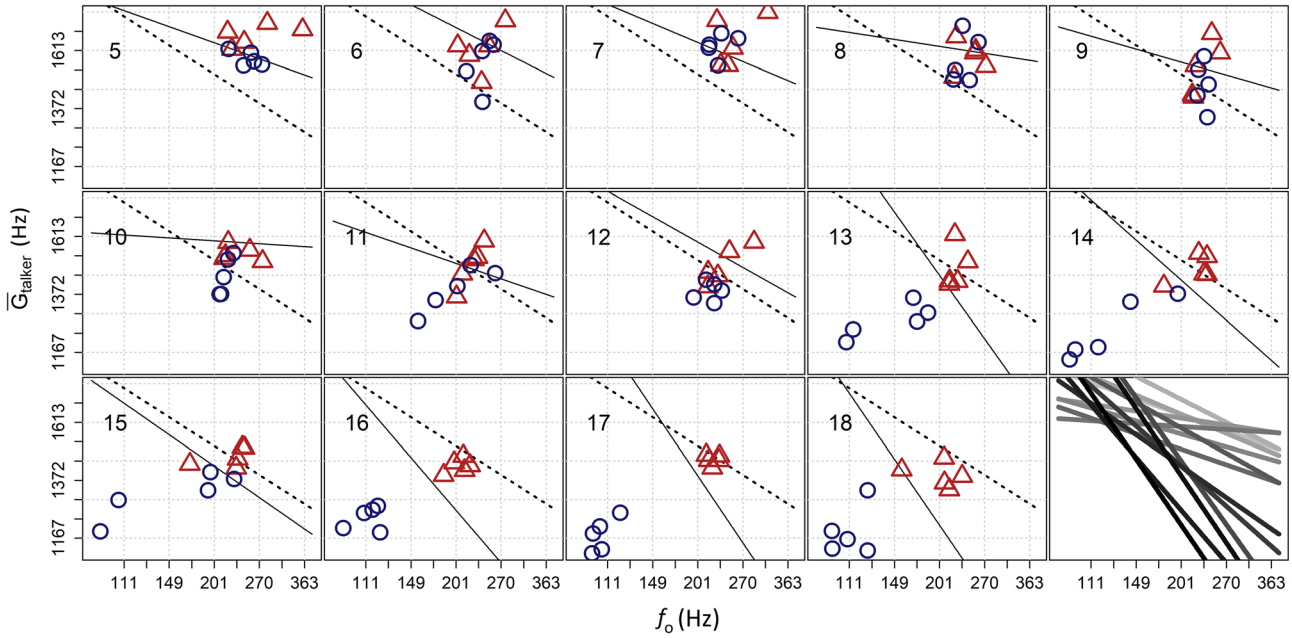

FIG. 10. (Color online) Each panel shows male (blue circles) and female (red triangles) talkers at different ages (indicated by numbers in each panel). Solid lines indicate boundaries based on the age-dependent intercept, f_o and $\overline{G}_{talker}$ estimated from our acoustic model. Dotted lines indicate boundaries based on the fixed use of the same predictors in the acoustic model. In the final panel (bottom right), all of the age-dependent boundaries are compared, with darker colors indicating older age groups.

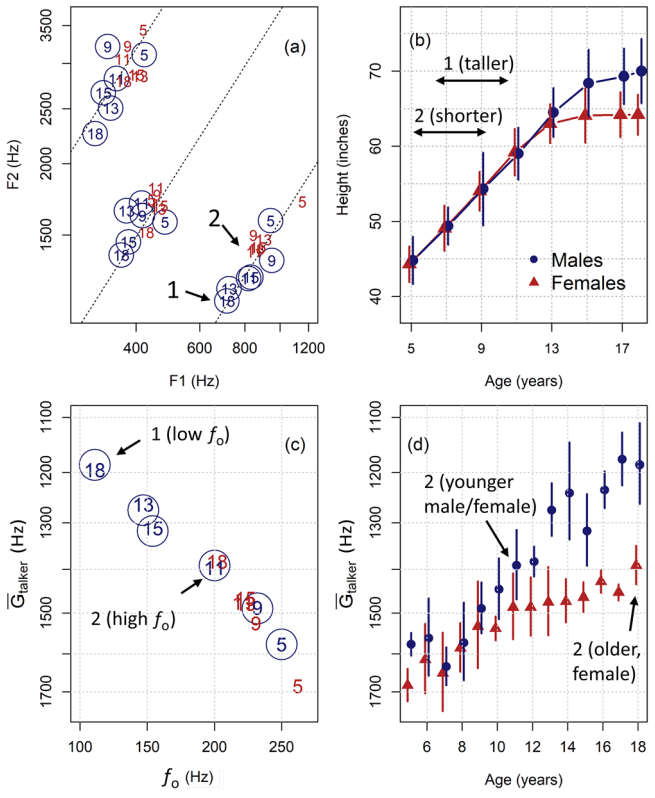

FIG. 11. (Color online) (a) Average F1 and F2 values for each vowel, for a subset of age and gender groups. Numbers indicate age groups (males in circles). (b) Average height for different age and gender groups (Fryar *et al.*, 2012). Error bars enclose one standard deviation. (c) Average $\overline{G}_{talker}$ and f_o for different age and gender groups. Numbers indicate age groups (males in circles). (d) Distribution of $\overline{G}_{talker}$ by age and gender in our sample. Error bars enclose one sd. In each plot, red is used for females and blue is used for males.

must also commit to an interpretation of the talker's vocal tract. For example, if the listener encounters a production at location (1) in the figure and concludes that this is an /ɑ/, then they *must* also commit to a long vocal-tract for the talker (and a low $\overline{G}_{talker}$). On the other hand, if a listener encounters a vowel at location (2) in the figure and also concludes that it is an /ɑ/, then they must also conclude that the talker has a shorter vocal tract, and produces higher formants overall (i.e., has a higher $\overline{G}_{talker}$).

Several indexical characteristics (e.g., age, sex, gender, height) correlate strongly with $\overline{G}_{talker}$, and are also highly correlated with each other. So, in addition to fixing an estimate of $\overline{G}_{talker}$ and approximate vocal-tract length, committing to an interpretation of a vowel in some cases may also affect the perception of talker indexical characteristics (or *vice versa*). Height correlates very strongly with vocal-tract length across the human population (Fitch and Giedd, 1999), and the age of children is almost perfectly correlated with their average height, especially before 15 years of age (Fryar *et al.*, 2012). As a result, locations 1 and 2 in Fig. 11(a) force associations with talkers of different heights, and potentially ages and genders, as seen in Fig. 11(b). In addition, the $\overline{G}_{talker}$ implied by locations 1 and 2 in Fig. 11(a) suggest different expectations about talker f_o [Fig. 11(c)], which typically has a strong association with talker gender. As a result of these relationships, if a listener gets an estimate of a talker's vocal-tract length for "free" in the process of vowel identification, they also get an estimate of the talker's approximate height and, in the case of children, the approximate age. Although the joint consideration of indexical characteristics and vocal-tract length may not necessarily solve all ambiguous cases, it may help by narrowing

down the range of possible solutions. For example, consider location 2 in Fig. 11(c), corresponding to either 11-year-old males or 18-year-old females. These talkers have similar f_o and $\overline{G}_{talker}$ values, meaning that these gross acoustic cues cannot be used to accurately identify their gender. As seen in Fig. 11(b), these talkers also roughly overlap in their approximate heights. However, in Fig. 11(d) we see that these talkers are separable along the age dimension: talkers are either younger males or older females, but not the other way around. As a result, anything that suggests "maturity" will help reduce ambiguity and increase the probability that both age and gender are identified correctly. More generally, cues that suggest either "femininity" *or* "maturity" can potentially affect the accurate perception of *both* talker age and gender when these cues are considered together.

D. The bias towards male responses

In Sec. III A, we discussed the bias towards responding "male" for post pubescent voices, a pattern of results which has been reported previously (e.g., Amir *et al.*, 2012; Lass *et al.*, 1976). Owren *et al.* (2007) suggest that the bias towards male responses may be because "adult male voices can be considered 'marked' by the sexually selected features of lowered f_o and formant frequencies," which "virtually guarantees that the talker is an adult male." However, as the authors note, "their absence does not unequivocally imply that the talker is an adult female. The individual might, for instance, be a young male whose voice has not yet changed, or a post pubertal male whose vocal tract has not diverged as far from the female form as is typically the case" (p. 931).

Figures 9 and 10 support the general interpretation of the response bias outlined previously. The increase in male bias corresponds approximately to the appearance of low-f_o voices in the male talkers. As seen in Fig. 11, low-f_o males are the most discriminable group, and these talkers can be unambiguously classified independently of age information. An increase in correct male responses for these talkers, and less accurate (but unbiased) responses for older girls (who are potentially confusable with younger males) would result in an increase in the false alarm rate (older girls confused with younger boys), without a corresponding increase in misses (older males identified as females). Thus, the bias towards male responses for older talkers may simply reflect the fact that the gender of older male talkers with low f_os is generally easy to identify.

The previous result should not be taken as indicating that the acoustic features most typically associated with males (e.g., a low f_o) are necessary to communicate maleness. In this experiment, we found the accurate perception of maleness in the complete and total absence of the acoustic features which are traditionally thought to "mark" a voice as male (low f_o and low formants). Thus, although the bias towards male responses may very well be because many adult male voices are "marked" by unique features, young boys are able to express maleness even in the absence of these features. This suggests that the communication of gender from speech does not necessarily rely on sexually selected anatomical differences between men and women, as these tend to emerge after puberty. Essentially, if the only way to hear "maleness" in a voice was to hear a low f_o and low formants, no pre-pubescent child should ever sound male.

E. Evidence for gendered speech patterns in the voices of children

Figure 6(c) indicates substantial variation in the by-talker random intercepts in our acoustic model. These effects capture the tendency of listeners to consistently identify specific talkers as male or female, above and beyond what can be explained by the acoustic predictors in our model (even considering the listener- and age-dependent adjustments). When these by-talker intercepts differ from zero in the 'correct' direction (i.e., negative intercepts for female talkers), these reflect 'residual' gender information in voices: the tendency for listeners to identify talker gender more accurately than can be explained by our acoustic model. As seen in Fig. 12 and 61% of intercepts associated with female talkers have negative posterior means, while 69% of intercepts for the male talkers have positive posterior means. This means that about two thirds of talker-intercepts represent talkers whose gender was identified more accurately than can be explained by our acoustic model.

In Sec. III A, we reported an advantage for gender identification in sentences over syllables. Although we did not

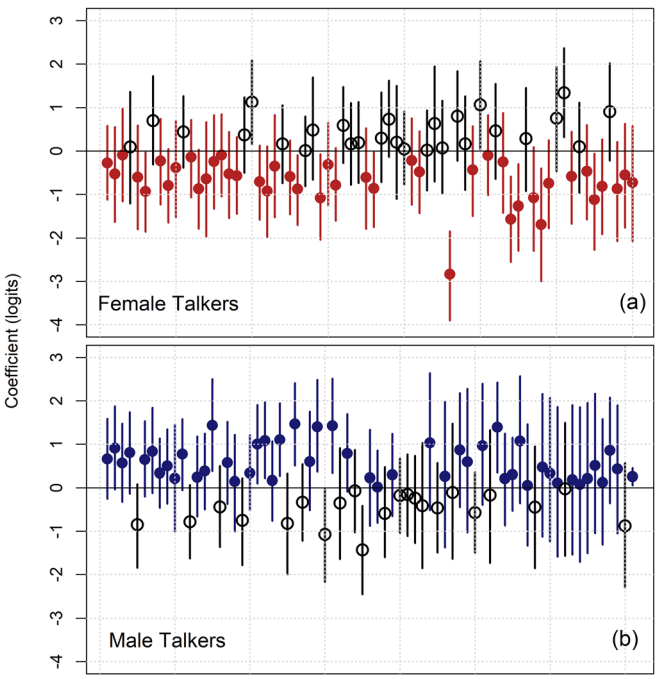

FIG. 12. (Color online) Distribution of talker random intercepts for (a) female and (b) male talkers, ordered by age within sex. Filled points indicate talker intercepts that reflect more accurate gender perception than can be explained by the model for that talker. Points indicate means, bars indicate 95% highest density intervals of posterior distributions.

https://doi.org/10.1121/10.0006785

investigate the nature of this effect acoustically in this study, our results suggest that the stimuli in the sentence and syllable conditions contained consistent gender information for most talkers. Figure 13(a) compares classification rates for individual talkers across both listening contexts. Talkers near the diagonal were classified in the same way in both conditions. Male talkers above the diagonal, and female talkers below the diagonal, were classified more accurately in the sentence context. Males in the top-left quadrant and females in the bottom-right quadrant represent cases where classification changed from incorrect in the syllable context, to correct in the sentence context. There are no notable cases of the opposite effect, where classifications change from correct in syllables to incorrect in sentence contexts.

The distribution of female talkers in Fig. 13(a) resembles a fan shape, with talkers spreading out more along the *y* axis as one moves left to right along the *x* axis. A similar distribution is seen for male voices, spreading out from the top right of the figure. These shapes indicate that in many cases where voices are somewhat ambiguous for isolated

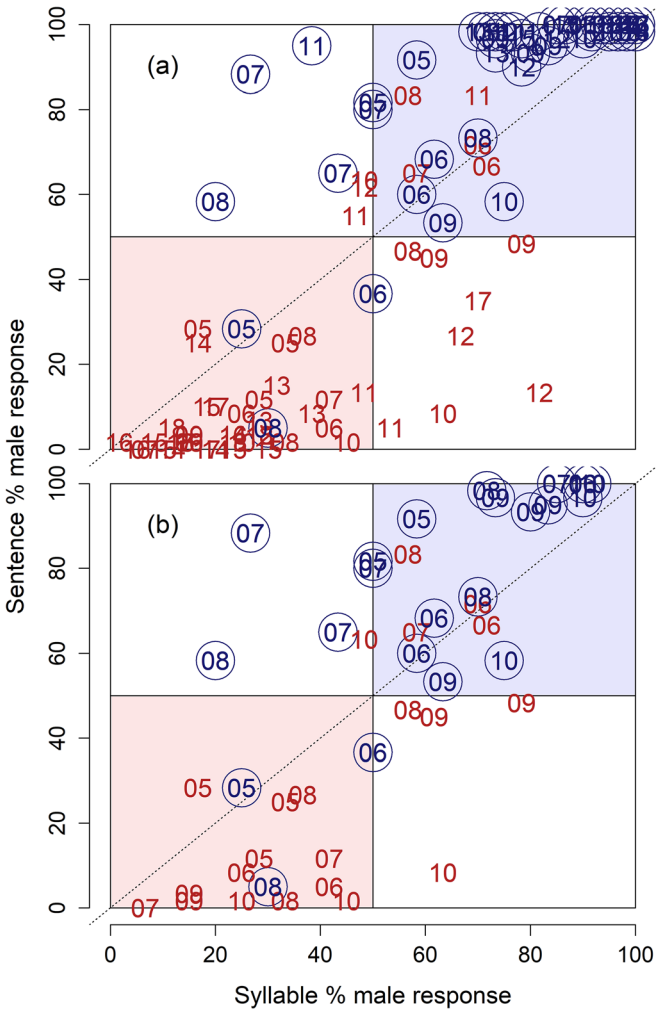

FIG. 13. (Color online) (a) Classification rates for individual talkers across both information conditions. Numbers indicate talker age (males in circles). (b) Only talkers under 11 years of age are presented. Voices in shaded quadrants were identified correctly in both conditions.

syllables, listeners are extremely accurate for the same voice in a sentence context. However, given the overlap of male and female talkers according to gross acoustic cues shown in Fig. 10(c), it seems unlikely that the sentence improvement could be solely attributed to a more precise estimation of the same, overlapping acoustic cues. This suggests that the key difference may be the presence of more dynamic, prosodic information in the sentence context (Hillenbrand and Clark, 2009). Since it seems unlikely that gender-based differences in prosodic structure would be explainable entirely on the basis of children's anatomy, this interpretation further supports the constructivist perspective of gender information in children's voices.

Finally, a careful consideration of Figs. 12 and 13 suggests that questions such as "can listeners identify the gender of 7-year olds?" are to some extent ill posed. Figure 13(b) shows that the gender of many voices can be correctly identified almost perfectly for talkers under 11 years of age, despite the fact that boys and girls of these ages substantially overlap in their f_o and formant frequencies [see Fig. 11(d)]. Since the expression of gender identity in voice occurs at the individual level, the gender of one talker may be correctly identified almost always, while the gender of another talker may be difficult to identify, or be consistently misidentified. Thus, although the classification of gender may be more difficult in prepubescent children, it is clear that gender information is being conveyed clearly in the voices of at least some of these children. In other words, the general confusability of gender in 7-year-old voices does not negate the clear expression of gender in any one seven-year-old talker. Taken together, our results provide support for the notion that the transmission of gender information from voice can depend on gender-dependent patterns of articulation, rather than following deterministically from anatomical differences between male and female talker (Zimman, 2018).

V. CONCLUSION

Our results indicate that children's gender can be accurately identified from their speech, in particular when listeners are presented with a longer stretch of speech (i.e., sentences vs syllables). Since young boys and girls overlap almost entirely in the gross acoustic cues that drive gender perception for adults (f_o, average formant frequencies), our results also suggest that this accuracy is based on the transmission of more subtle gender information. Two likely candidates for the "more subtle" acoustic information are prosodic cues and the source characteristics. Although we did not find any important role for the source-related predictors (apart from f_o), we do not think that this demonstrates that these predictors have no role in gender perception. Instead, as with the role for prosodic information, it may be the case that source information plays an important, but potentially complicated role in gender perception from children's voices. Overall, the accuracy of gender identification for young boys and girls in the absence of reliable

anatomical differences between these talkers supports the constructivist view that voice gender information has a strong performative component, rather than following necessarily from talker anatomy.

ACKNOWLEDGMENTS

This research was supported by a grant from the National Science Foundation (Grant No. 1124479, P.F.A). We thank Terry Nearey, Daniel Hubbard, and Michelle Kapolowicz for their helpful comments and discussion.

[1]If talkers A and B produce approximately the same articulatory gesture but differ in vocal tract length by a factor of x, A will produce formant frequencies that are $1/x$ lower than B (Wakita, 1977). Such a difference in formant scaling would manifest as a $log(1/x)$ difference between $\overline{G}_{talker}$ estimates for talkers A and B.

[2]This approach can be implemented using the regression approach to normalization outlined in Barreda and Nearey (2018). When a regression model is used to carry out log-mean normalization as described in Barreda and Nearey, the estimated talker main effects are the least-squares estimates of Gtoken under a fairly broad set of conditions.

Amir, O., Engel, M., Shabtai, E., and Amir, N. (**2012**). "Identification of children's gender and age by listeners," J. Voice **26**(3), 313–321.
Assmann, P., Ballard, W., Bornstein, L., and Paschall, D. (**1994**). "TrackDraw: A graphical interface for controlling the parameters of a speech synthesizer," Behavior Res. Methods Instrum. Comput. **26**, 431–436.
Assmann, P. F., Nearey, T. M., and Bharadwaj, S. (**2008**). "Analysis of a vowel database," Can. Acoust. **36**(3), 148–149.
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**(2), 1054–1063.
Barreda, S. (**2017**). "An investigation of the systematic use of spectral information in the determination of apparent-talker height," J. Acoust. Soc. Am. **141**(6), 4781–4792.
Barreda, S. (**2020**). "Vowel normalization as perceptual constancy," Language **96**, 224–254.
Barreda, S., and Assmann, P. F. (**2018**). "Modeling the perception of children's age from speech acoustics," J. Acoust. Soc. Am. **143**(5), EL361–EL366.
Barreda, S., and Nearey, T. M. (**2018**). "A regression approach to vowel normalization for missing and unbalanced data," J. Acoust. Soc. Am. **144**(1), 500.
Bürkner, P.-C. (**2018**). "Advanced Bayesian multilevel modeling with the R package brms," R. Journal **10**(1), 395–411.
Cartei, V., Banerjee, R., Hardouin, L., and Reby, D. (**2019a**). "The role of sex-related voice variation in children's gender-role stereotype attributions," Br. J. Dev. Psychol. **37**(3), 396–409.
Cartei, V., Garnham, A., Oakhill, J., Banerjee, R., Roberts, L., and Reby, D. (**2019b**). "Children can control the expression of masculinity and femininity through the voice," R. Soc. Open Sci. **6**(7), 190656.
Cartei, V., and Reby, D. (**2013**). "Effect of formant frequency spacing on perceived gender in pre-pubertal children's voices," PLoS One **8**(12), e81022.
Clopper, C. G., and Smiljanic, R. (**2011**). "Effects of gender and regional dialect on prosodic patterns in American English," J. Phon. **39**(2), 237–245.
DeCarlo, L. T. (**1998**). "Signal detection theory and generalized linear models," Psychol. Methods **3**(2), 186–205.
de Krom, G. D. (**1993**). "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," J. Speech Lang. Hear. Res. **36**(2), 254–266.
Fitch, W. T., and Geidd, J. (**1999**). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," J. Acoust. Soc. Am. **106**(3), 1511–1522.
Fryar, C. D., Gu, Q., and Ogden, C. L. (**2012**). "Anthropometric reference data for children and adults: United States, 2007–2010," Vital Health Stat. 11 **252**, 1–48, PMID: 25204692.
Gelman, A. (**2005**). "Analysis of variance—Why it is more important than ever," Ann. Stat. **33**(1), 1–53.
Gelman, A., Hill, J., and Yajima, M. (**2012**). "Why we (usually) don't have to worry about multiple comparisons," J. Res. Educat. Effect. **5**(2), 189–211.
Gerosa, M., Giuliani, D., and Brugnara, F. (**2007**). "Acoustic variability and automatic recognition of children's speech," Speech Commun. **49**(10), 847–860.
Hillenbrand, J., Cleveland, R. A., and Erickson, R. L. (**1994**). "Acoustic correlates of breathy vocal quality," J. Speech Hear. Res. **37**(4), 769–778.
Hillenbrand, J. M., and Clark, M. J. (**2009**). "The role of f_o and formant frequencies in distinguishing the voices of men and women," Atten. Percept. Psychophys. **71**(5), 1150–1166.
Holway, A. H., and Boring, E. G. (**1941**). "Determinants of apparent visual size with distance variant," Am. J. Psychol. **54**(1), 21–37.
Ingemann, F. (**1968**). "Identification of the speaker's sex from voiceless fricatives," J. Acoust. Soc. Am. **44**(4), 1142–1144.
Iseli, M., Shue, Y.-L., and Alwan, A. (**2007**). "Age, sex, and vowel dependencies of acoustic measures related to the voice source," J. Acoust. Soc. Am. **121**(4), 2283–2295.
Johnson, K. (**2020**). "The ΔF method of vocal tract length normalization for vowels," Lab. Phonol. **11**(1), 10.
Kawahara, H., de Cheveigné, A., Banno, H., Takahashi, T., and Irino, T. (**2005**). "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," in *Proc*eedings of *Interspeech 2005*, September 4–8, Lisboa, Portugal, pp. 537–540.
Khwaileh, F. (**2011**). "Temporal and aerodynamic aspects of velopharyngeal coarticulation: Effects of age, gender and vowel height," Ph.D. thesis, University of Tennessee, Knoxville, TN.
Lammert, A. C., and Narayanan, S. S. (**2015**). "On short-time estimation of vocal tract length from formant frequencies," PLoS ONE **10**(7), e0132193.
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**(3), 675–678.
Lee, S., Potamianos, A., and Narayanan, S. (**1999**). "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," J. Acoust. Soc. Am. **105**(3), 1455–1468.
Munson, B., and Babel, M. (**2019**). "The phonetics of sex and gender," in *The Routledge Handbook of Phonetics* (Taylor and Francis, London), pp. 499–525.
Nearey, T. M. (**1978**). *Phonetic Feature Systems for Vowels* (Indiana University Linguistics Club, Bloomington, IN).
Nearey, T. M., and Assmann, P. F. (**2007**). "Probabilistic 'sliding-template' models for indirect vowel normalization," in *Experimental Approaches to Phonology*, edited by M.-J. Solé, P. S. Beddor, and M. Ohala (Oxford University Press, Oxford), pp. 246–269.
Nearey, T. M., Assmann, P. F., and Hillenbrand, J. M. (**2002**). "Evaluation of a strategy for automatic formant tracking," J. Acoust. Soc. Am. **112**, 2323.
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**(6), 930–941.
Perry, T. L., Ohde, R. N., and Ashmead, D. H. (**2001**). "The acoustic bases for gender identification from children's voices," J. Acoust. Soc. Am. **109**(6), 2988–2998.
R Core Team (**2019**). "R: A Language and Environment for Statistical Computing," http://www.R-project.org (Last viewed October 26, 2021).
Sachs, J., Lieberman, P., and Erickson, D. (**1973**). "Anatomical and cultural determinants of male and female speech," in *Language Attitudes: Current Trends and Prospects*, edited by R. W. Shuy and R. W. Fasold (Georgetown University Press, Washington, D.C.).
Schwartz, M. F. (**1968**). "Identification of speaker sex from isolated, voiceless fricatives," J. Acoust. Soc. Am. **43**(5), 1178–1179.
Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. (**2009**). "Voicesauce. P. Program," http://www.Seas.Ucla.Edu/Spapl/Voicesauce/UCLA (Last viewed October 26, 2021).
Smith, D. R. (**2016**). "Speaker-sex discrimination for voiced and whispered vowels at short durations," I-Perception **7**(5), 204166951667132.
Story, B. H., Vorperian, H. K., Bunton, K., and Durtschi, R. B. (**2018**). "An age-dependent vocal tract model for males and females based on anatomic measurements," J. Acoust. Soc. Am. **143**(5), 3079–3102.
Turner, R. E., Walters, T. C., Monaghan, J. J. M., and Patterson, R. D. (**2009**). "A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data," J. Acoust. Soc. Am. **125**(4), 2374–2386.

Vorperian, H. K., Wang, S., Chung, M. K., Schimek, E. M., Durtschi, R. B., Kent, R. D., Ziegert, A. J., and Gentry, L. R. (**2009**). "Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study," J. Acoust. Soc. Am. **125**(3), 1666–1678.

Vorperian, H. K., Wang, S., Schimek, E. M., Durtschi, R. B., Kent, R. D., Gentry, L. R., and Chung, M. K. (**2011**). "Developmental sexual dimorphism of the oral and pharyngeal portions of the vocal tract: An imaging study," J. Speech Lang. Hear. Res. **54**(4), 995–1010.

Wakita, H. (**1977**). "Normalization of vowels by vocal-tract length and its application to vowel identification," IEEE Trans. Acoust. Speech Signal Process. **25**(2), 183–192.

Weinberg, B., and Bennett, S. (**1971**). "Speaker sex recognition of 5- and 6-year-old children's voices," J. Acoust. Soc. Am. **50**(4B), 1210–1213.

Zimman, L. (**2018**). "Transgender voices: Insights on identity, embodiment, and the gender of the voice," Lang. Linguist. Compass **12**(8), e12284.