Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex

Jason Bishop[a] and Patricia Keating
Department of Linguistics, University of California, 3125 Campbell Hall, Box 951543, Los Angeles, California 90095-1543

(Received 22 July 2010; revised 13 March 2012; accepted 19 April 2012)

How are listeners able to identify whether the pitch of a brief isolated sample of an unknown voice is high or low in the overall pitch range of that speaker? Does the speaker's voice quality convey crucial information about pitch level? Results and statistical models of two experiments that provide answers to these questions are presented. First, listeners rated the pitch levels of vowels taken over the full pitch ranges of male and female speakers. The absolute f0 of the samples was by far the most important determinant of listeners' ratings, but with some effect of the sex of the speaker. Acoustic measures of voice quality had only a very small effect on these ratings. This result suggests that listeners have expectations about f0s for average speakers of each sex, and judge voice samples against such expectations. Second, listeners judged speaker sex for the same speech samples. Again, absolute f0 was the most important determinant of listeners' judgments, but now voice quality measures also played a role. Thus it seems that pitch level judgments depend on voice quality mostly indirectly, through its information about sex. Absolute f0 is the most important information for deciding both pitch level and speaker sex.
 [http://dx.doi.org/10.1121/1.4714351]

PACS number(s): 43.71.Bp, 43.71.Hw [PEI] Pages: 1100–1112

I. INTRODUCTION

Fundamental frequency (f0) conveys linguistic (e.g., tone, intonation), paralinguistic (e.g., emotion, emphasis), and non-linguistic (e.g., physiology) information from a speaker, but this information is relative to what is a low or high f0 for that speaker. Thus listeners' interpretation of the speaker's use of f0 depends crucially on their ability to infer a speaker's overall f0 range and thereby normalize individual f0 values. An important question is *how* listeners are able to do this.

Honorof and Whalen (2005) showed that listeners could judge the locations of very brief voice samples in speakers' f0 ranges. Their stimuli were 500 ms isolated steady-state vowels by 20 English speakers, produced with f0s throughout the speakers' ranges (as determined in a separate production task). Lee (2009), building on previous work on tone identification, then showed that Mandarin listeners could distinguish tones that begin with relatively high f0s (the high-level Tone 1 and the high-falling Tone 4) from tones that begin with relatively low f0s (the mid-rising Tone 2 and the low-dipping Tone 3). His stimuli were the fricative plus the first six glottal pulses of the vowel taken from /sa/ syllables, produced by 32 Mandarin speakers.

By design, these experiments eliminated some potential sources of information for listeners. Because the stimuli were from unknown speakers and isolated from any context, and the f0s were virtually steady, factors such as familiarity or experience with an individual speaker, sentential context, dynamic f0, or specific intonation contours (Leather, 1983; Wong and Diehl, 2003; Moore and Jongman, 1997; Greenberg and Zee, 1979) can be ruled out. The implication, then, is that listeners must use other signal-internal information as cues to a speaker's f0 range. Both Honorof and Whalen (2005) and Lee (2009) speculated that voice quality could be such a cue, and/or that listeners could have identified the sex of the speakers, and then made sex-specific decisions about f0 locations by applying experience-based knowledge of sex-specific f0 ranges (population ranges stored in memory). Indeed Lee *et al*. (2010) and Honorof and Whalen (2010) showed that listeners could identify the speaker sex of the tokens used in their respective earlier studies. In this paper we pursue the relation of f0 and of voice quality to judgments of location-in-f0-range and of speaker sex.

A. Voice quality as a cue to location-in-f0-range and to speaker sex

There are at least three ways in which voice properties could be related to f0. First, there could be a *direct* relation to the f0 *range*, meaning that a voice measure has a consistent relation to relative f0, in the same way for every individual speaker. For example, every individual speaker would have a low value on some voice measure for his or her own low f0s, and a high value for his or her own high f0s, regardless of the absolute f0s. In this case, the voice measure would give listeners clear and independent information about the location of any given f0 in that speaker's f0 range. Swerts and Veldhuis (2001) showed that f0 correlated with H1-H2 for individual male Dutch speakers; but these within-speaker correlations were very variable, indeed in opposing directions, across the speakers, and thus provide little support for this first scenario.

Second, there could be a *direct* relation to the *absolute* f0, meaning that across all speakers a voice measure has a

[a]Author to whom correspondence should be addressed. Electronic mail: j.bishop@ucla.edu

 0001-4966/2012/132(2)/1100/13/$30.00

consistent relation to f0, regardless of each speaker's f0 range. For example, speakers who speak with low f0s would have low values on some measure, while speakers who speak with high f0s would have high values. In this case, the voice measure would provide redundant information about the f0, but no information about the location of that f0 in any individual speaker's range. Such a correlation across speakers has been found between f0 and H1*-H2* (where the asterisks indicate corrections for formant frequencies and bandwidths) by Iseli *et al.* (2007); and Lee (2009) showed modest correlations of f0 with H1-A1 and H1-A3 (uncorrected) for a sample that was mostly across-speaker. Thus there is evidence that at least these voice measures likely do *not* provide listeners with information about a speaker's f0 range independently of f0 itself. Indeed, Lee also found in a regression analysis that none of the voice measures H1-A1, H1-A3, or H1-H2 contributed significantly to listeners' tone identifications.

The last way in which voice properties could be related to f0 would be *indirect*, where voice quality serves as a cue to some other speaker characteristic that allows the listener to interpret f0 information, e.g., by referring to an f0 range already stored in memory for such speakers. As noted by previous researchers, including Honorof and Whalen (2005) and Lee (2009), speaker sex is such a characteristic. There are a number of studies that indicate that the speech signal differs between the sexes in ways that could be exploited in perception. A large literature [see Kreiman and Sidtis (2011) and references therein] suggests that identification of male and female voices is very well predicted by formant frequencies and f0, listeners being biased towards hearing a female voice when these are above the average male values. But voice properties could also play a role: female voices have higher H1-H2 (Henton and Bladon, 1985; Klatt and Klatt, 1990), H1-A3 (Perkell *et al.*, 1994), and H1-A1 (Hanson and Chuang, 1999). Indeed, Shue (2010) found that voice measures can improve automatic sex classification, though significantly so only for 10-to-14-year-old children's voices. Similarly, Mecke and Sundberg (2010) found that listeners' judgments of the sex of 10-to-13-year-old singing children correlated well with the Closed Quotient from electroglottographic signals. However, vowel F4 correlated best, and F2, F3 and F5 were also correlated with the responses, suggesting that listeners use vocal tract information more than voice quality information in this decision. Both Honorof and Whalen (2010) and Lee *et al.* (2010) hypothesized that voice quality could be a cue for speaker sex. However, Honorof and Whalen compared different cues for sex identification, and concluded that their listeners relied on sex-specific f0 and perhaps formant values, but gave "no strong evidence for a contribution of voice quality" (p. 3095), represented in their study by the voice measures jitter and shimmer. Possibly, then, voice quality does not contribute even indirectly to f0 judgments; here we further test this hypothesis.

B. Present study

The two experiments below explore what factors contribute to a listener's placement of an f0 in the ranges of individual speakers, for brief stimuli as in Honorof and Whalen's (2005) study. In the present study, we are especially interested in whether voice quality plays a role, and whether listeners have language-specific strategies for performing the task. English and Mandarin serve as our target languages. Experiment 1 replicates the basic findings reported in Honorof and Whalen (2005). We then build a model of listeners' perception based on a range of possible cues in the signal, including information about speaker sex. In experiment 2 we investigate, also by way of a model, what cues listeners use to distinguish the sex of the speakers, using a large set of parameters similar to that in experiment 1. To establish the contribution of voice quality, it is necessary to show that it independently accounts for some of the variance in listeners' judgments.

II. EXPERIMENT 1

A. Method

1. Stimuli

a. Speakers. Ten adult native speakers of English and 10 of Mandarin (5 males and 5 females of each language) participated in a production task. English speakers came from diverse locations throughout the US, and Mandarin speakers came from either mainland China or Taiwan (in neither case was specific dialectal information retained). All Mandarin speakers spoke English, although with varying degrees of proficiency. All speakers confirmed that they did not have any known speech, hearing or communication disorders. There was no screening for a history of smoking, nor for formal training in music or singing.

b. Design and creation of stimuli. In order to create brief f0 samples from various locations in speakers' ranges (to be presented to listeners later), speakers were asked to perform two tasks. The first was designed to estimate the speakers' individual f0 ranges, using a common method for clinical or experimental purposes whereby speakers produce rising or falling spoken glissandos using the vowel /ɑ/ (Reich *et al.*, 1990; Zraick *et al.*, 2000, Honorof and Whalen, 2005, and supplementary material[1]). Recordings took place in a sound-attenuated booth using a Shure SM10A head-mounted microphone (50–15 000 Hz frequency response), digitized at 44.1 kHz and 24 bits by an XAudio A/D box with PCQuirer, and later converted to WAV files. Speakers' individual f0 ranges (highest and lowest sustainable f0) were determined via inspection of f0 tracks made by the autocorrelation method in Praat (Boersma and Weenink, 2008). The most extreme high and low was identified for each speaker, and these are shown as dashed lines in Fig. 1. Even though all the speakers were speaking, not singing, almost every speaker's high pitches were produced in falsetto register (sometimes called "loft" in the speech literature). Unlike in Honorof and Whalen's study, these pitches were retained, for three reasons. First, standard definitions of, and data for, the Maximal Phonational Frequency Range include falsetto (e.g., Baken and Orlikoff, 2000), with maximum f0s in some studies well above 1100 Hz, even for men. Second, there is

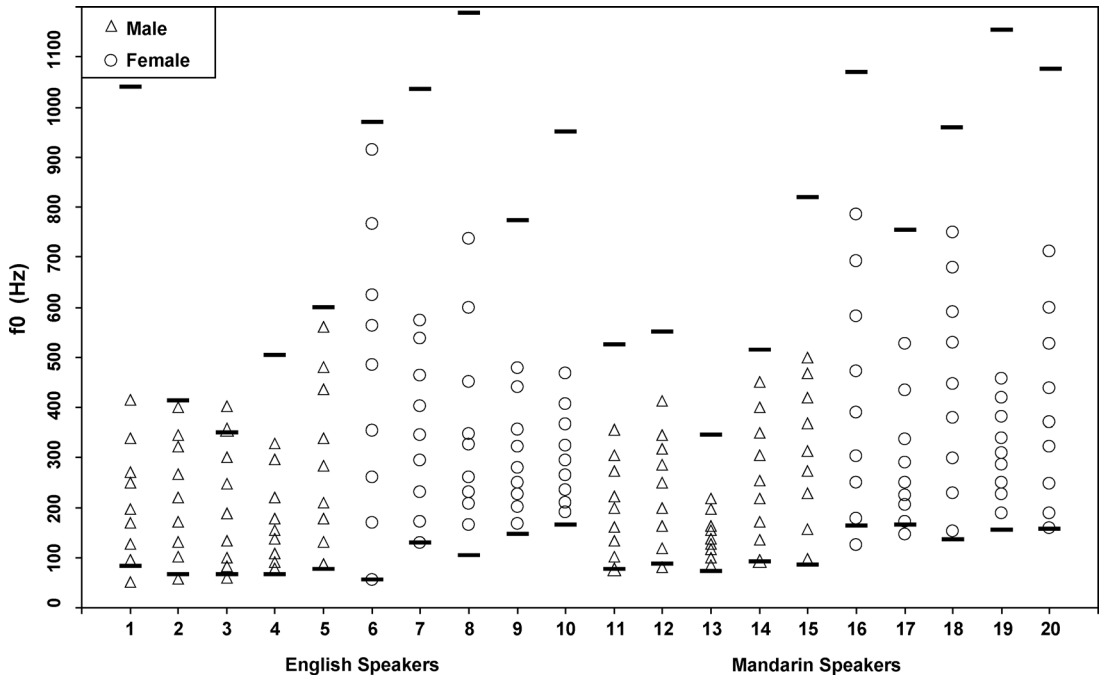

FIG. 1. F0s of tokens from English (1–10) and Mandarin (11–20) speakers selected from the step task for use as stimuli in the perception experiment. The horizontal dashes at the low and high ends of each speaker's pitch range represent the f0 extremes produced in the separate sweep task.

no completely reliable basis for determining whether any pitch is in falsetto register and thus should be excluded. Finally, listeners' expectations about whether falsetto pitches are part of a speaker's possible pitch range are unknown. Use of falsetto in speaking is not uncommon among some English speakers, whether humorously or linked to sociolinguistic factors (e.g., Podesva, 2007), so listeners might reasonably expect a speaker's high pitches to be falsetto.

In a second production task, speakers were asked to produce brief, steady-state /ɑ/ vowel tokens. This task closely resembled the task used to elicit the steady-state tokens in Honorof and Whalen's (2005) study, although unlike those authors, we did not prompt speakers to produce tokens within their already-established glissando range. Rather, speakers produced the level tokens (maintained for approximately 3–4 s each), in f0 steps spanning as high and as low in their range as possible. These steady-state recordings were then used to create the stimuli. 500 ms portions were extracted from the tokens most free of f0 or amplitude excursions or perturbations, and a 50 ms linear amplitude ramp was applied to the beginning and end of each. The highest- and lowest-f0 tokens recorded were then selected, and the seven tokens most evenly spaced between those two f0s were selected from the much larger available set. This resulted in 180 tokens (9 tokens × 20 speakers).[1]

The f0s for these tokens, and their locations in each speaker's range, are shown in Fig. 1. As can be seen there, for most speakers the tokens elicited in the step task were in fact a subset of the speaker's range as determined in the glissando task [see also Reich *et al*. (1990) and Baken and Orlikoff (2000)]. Even so, for each speaker a substantial portion of their physiological range (and a much wider range than normal speaking f0), in most cases including falsetto register, was represented by the nine roughly equally spaced tokens.

Finally, in addition to the stimuli created from the twenty speakers' productions, a set of synthetic tone stimuli was created, the purpose of which was to investigate listeners' use of absolute f0 by itself. Synthetic tones do not sound like a human voice, and thus not only were no other acoustic properties of a voice available, but also no inferences could be made about a possible overall f0 range for an individual speaker. Thus the tone stimuli provide a kind of baseline measure of whether listeners have expectations about the f0 ranges of the population of human voices. Specifically, ten level sawtooth tones were created in Audacity (Various Contributors, 2008), ranging, in 50 Hz intervals, from 50 Hz to 950 Hz. Like the speech stimuli, these stimuli were 500 ms in duration and were given linear amplitude ramps of 50 ms at onsets and offsets.

c. Acoustic properties of the stimuli. A number of acoustic measures were collected for all 180 voice (not synthetic) tokens, with the mean of each measure taken over each entire stimulus. The frequencies of each of the first three formants (F1, F2, F3) were estimated in Praat, with careful manual adjustment of parameters to get the best estimates despite the very high f0s of some stimuli. These formant frequencies were then ported to the program VoiceSauce (Shue *et al*., 2011), which automatically collected several measures reflecting characteristics of voice (see, e.g., Blankenship, 2002, for a review). These included cepstral peak prominence (CPP, Hillenbrand *et al*., 1994) and the relative amplitudes of the first and second harmonic (H1*-H2*); both measures relate to perceived breathiness in linguistic (Esposito, 2010) and nonlinguistic (Klatt and Klatt, 1990; Hillenbrand *et al*., 1994; Hillenbrand and Houde, 1996) tasks. Also collected were measures

of the amplitudes of H1 relative to the first and third formants (H1*-A1* and H1*-A3*, respectively), and of H2 relative to H4 (H2*-H4*). Both H1*-A1* and H1*-A3* are measures of spectral tilt, H1*-A1* also reflecting the bandwidth of F1 (Hanson, 1997; Hanson and Chuang, 1999) and H1*-A3* perhaps reflecting the speed, abruptness, and/or simultaneity of vocal fold closure (Stevens and Hanson, 1995; Hanson, 1997). Where a measure includes a harmonic amplitude, an asterisk represents corrections made for formant frequency and estimated bandwidth, using the extension of Iseli *et al.* (2007) of Hanson's (1997) proposal. These corrections give harmonic amplitudes closer to those in the voice source, unmodified by vocal tract resonances. Because of the robustness of VoiceSauce and its algorithms, and the careful checking of formant frequency estimates here, the measurements are believed to be reliable across the wide range of f0s in this stimulus set.

H2*-H4* is not a commonly used measure in studies of voice quality, and so deserves more comment. H2-H4 was introduced without explanation or citation in Kreiman *et al.* (2007), a comparison of 70 voice samples on a wide variety of voice measures. Principal components analysis of 19 measures from the spectrum of the full audio signal indicated that four components (associated with H1-H2, overall spectral slope, high-frequency noise excitation, and H2-H4) accounted for 76.6% of the variance in the measures, with H2-H4 accounting for 8.3%. Thus H2-H4 captures some important aspect of individual voice quality that is distinct from other, more familiar measures. Exploratory work in our lab suggests that voice samples produced by John Laver for the recordings accompanying Laver (1980) and characterized as varieties of "falsetto" are distinguished primarily by their low values on H2*–H4*. While contrastive phonation types in languages have not been found to differ on this measure (Keating *et al.*, 2011; Kuang, 2011), Kreiman and Garellek (2011) recently showed that Hmong listeners use higher values of source H2-H4 as a cue to their contrastive breathy phonation. Furthermore, Zhang *et al.* (2011), using a physical model of the vocal folds, found that higher values of source H2-H4 are associated with reduced stiffness of the body-layer of the folds. In sum, high H2*-H4* seems to be associated with a less stiff vocal fold body-layer and/or breathy voice, and low H2*-H4* with a stiffer body-layer and/or falsetto phonation.[2]

2. Listeners

Twenty native speakers of American English (mostly from California) and 21 native speakers of Mandarin (either Mainland or Taiwanese) participated as listeners in a location-in-f0-range rating task. None had participated as speakers in the production task described above. The Mandarin-speaking listeners were bilingual in English to varying degrees. All participants confirmed that they were given no previous diagnosis of a communication disorder and, to the best of their knowledge, had normal hearing.

3. Procedure

Listeners were presented with the steady-state /ɑ/ tokens taken from nine points at different locations in speakers' f0 ranges. The stimuli were presented in two blocks, one with all 90 tokens of English voices and one with all 90 tokens of Mandarin voices, with the order of the two language blocks counterbalanced. Tokens within each language block were randomized differently for each listener. Listeners were told that they would hear "voices," but were not explicitly told the language of the speakers, or that voices from two different languages would be presented, or even that there would be multiple tokens from any one speaker. Stimuli were presented to listeners at a comfortable listening volume (held constant across listeners) over Sony MDR V500 closed, dynamic headphones (10–25 000 Hz frequency response) which were connected to a soundcard external to the computer presenting the stimuli. Participants were asked to listen to each of the voice stimuli and decide how high or low the pitch of a given token was in that particular speaker's own range. Specifically, listeners were told to consider for each token how much higher or how much lower in pitch that speaker could have produced the vowel, and to identify where the token fell in that range.

For the synthetic stimuli (presented after all the voices), listeners were told: "For the last part of the experiment, you will hear computerized tones instead of human voices. Think of how high and low human voices are, and then rate each tone as if it were a voice. If it were a voice, could it go higher? Could it go lower?" Listeners expressed no difficulty in performing this task. They may have drawn on their experience with the set of human voices just heard in the experiment, but the instructions did not direct them to do so.

Listeners' responses to all stimuli (human and synthetic) were collected using a MATLAB script that provided a graphical user interface with a button allowing them to play the token (as many times as they wished, although they were discouraged from listening more than three times), and a bar that allowed them, after listening, to slide an icon along a horizontal continuum. This bar coded the icon's location on a scale from 0 to 100. However, participants saw only the bar, not the numerical scale. They were told that this bar represented the speaker's pitch range, and were instructed that for each token they should slide the icon to the position in that range that the token came from. The left edge of the continuum (labeled "lowest") represented the very lowest pitch the speaker could produce, and the right edge (labeled "highest") represented the very highest pitch the speaker could produce. Instructions were given to speakers in their native language by a native-speaking English or Mandarin experimental assistant. Participants were given a practice trial with three extra-experimental voices, and exhibited no difficulties in performing the task or using the interface.

B. Results

1. Correlations with location-in-f0-range and with f0

Listeners' ratings were pooled for each of the tokens, and these averaged ratings were then plotted separately against two independent variables: (a) the location of the token in the speaker's individual f0 range, expressed as a proportion along the total range, and (b) the f0 of the token (in Hz). Both of these correlations are shown in Fig. 2.

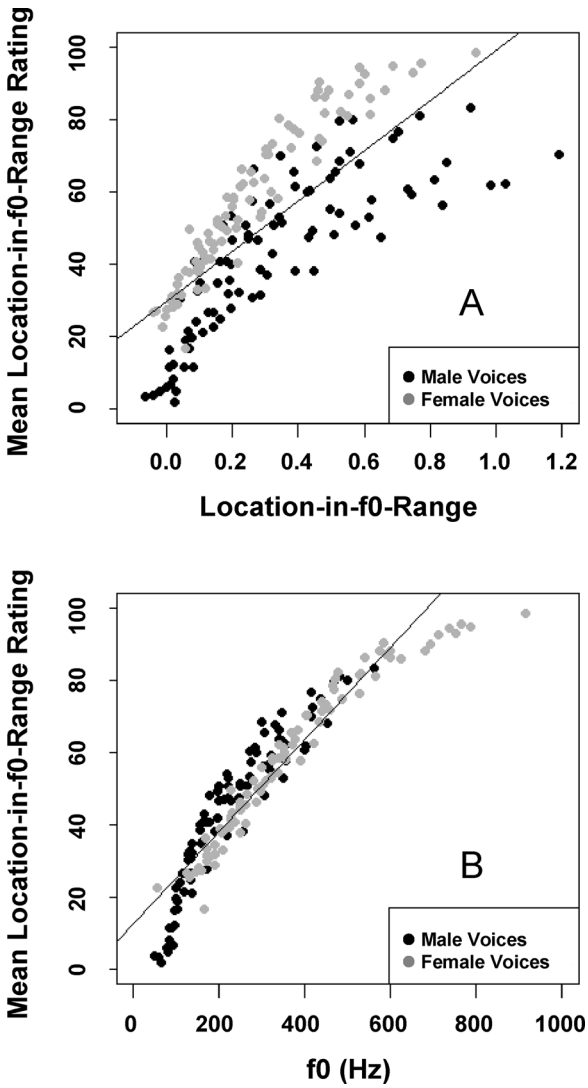

FIG. 2. Averaged (pooled across listeners) ratings of location-in-f0-range (A) as a function of the location in f0-range of the tokens (R^2=0.573) and (B) as a function of the absolute f0 of the token (R^2=0.88). Locations-in-f0-range are expressed as a proportion of speakers' individual range in the sweep task, and can be above 1.0 because the ranges calculated from that task differ from the ranges of the stimuli. Lines (and corresponding R^2s) are fit to the group as a whole (i.e., both sexes combined).

Regression lines were then fit to these data and are also shown in Fig. 2; these are always linear, even when a better fit could be found with another function.

Considering first the correlation with the actual location of the token in speakers' ranges [Fig. 2(a)], the best fit line indicates that the location-in-f0-range of the tokens accounts for 57.3% of the variance in the averaged listener ratings of the tokens (R^2=0.573). R^2 for the 41 individual listeners range from 0.076 to 0.601, generally similar to the values Honorof and Whalen (2005) reported, though some of our listeners did worse than theirs.[3] Unlike in Honorof and Whalen, however, here the correlation between location-in-f0-range and listeners' judgments of location- in-f0-range is somewhat stronger when the sexes are considered individually (R^2=0.655 for male voices, R^2=0.884 for female voices). A particularly interesting pattern can also be seen with respect to how the sexes were rated relative to one another: in general, a token at a given location in a speaker's f0 range was rated as being higher in the range if it came from a female speaker than if it came from a male speaker. (For example, in Fig. 2(a), it can be seen that tokens at about the 50% point in males' ranges were rated at about 50%, while tokens at about 50% in females' ranges were rated at about 80%.) It is somewhat puzzling why an f0 should be judged as coming from a higher location in a speaker's range simply by virtue of coming from a female speaker's voice, but the pattern is better understood when listeners' average judgments are plotted against absolute f0 rather than location-in-f0-range.

As can be seen in Fig. 2(b), listeners' judgments of location-in-f0-range show a much stronger relationship with absolute f0, which accounts for 88% of the variance in the averaged listener ratings of the tokens (R^2=0.88). The relationship was similar when the sexes were considered separately (R^2=0.942 for male voices, R^2=0.927 for female voices). This indicates that the absolute f0 was a better predictor of listeners' ratings than the actual location-in-f0-range was. Nonetheless, absolute f0 was apparently interpreted relative to the sex of a speaker: a male token at a given f0 was rated higher by listeners than a female token at the same f0. This is expected if listeners know that, on average, any f0 should be somewhat higher in the range of a male than a female, given the difference between male and female speaker f0 ranges. Likewise, a token at any given location in the range of a female should have a higher f0 than a token at the same location in a male range. As just shown, that was in fact the pattern.

2. Modeling listeners' judgments of location-in-f0-range

The correlations suggest that, at least when listeners are considered as a group, most of the variance is accounted for by f0. Nonetheless, it may be that factors such as voice quality contribute significantly to the remaining variance. To determine which of many possible acoustic properties of the stimuli could have influenced listeners' responses, those responses were modeled using mixed-effects linear regression (Baayen, 2008). In particular, we modeled the outcome "rating" (which, as described earlier, was a value from 0 to 100), using speaker and item as random intercept effects, and the following fixed-effects factors: (a) the language of the listener (English or Mandarin); (b) speaker sex (male or female); (c) the f0 of the token; (d) the mean frequency of each of the first three formants (F1, F2, F3); (e) mean measures of voice quality: CPP, H1*-A1*, H1*-A3*, H1*-H2*, H2*-H4*. In addition, 18 interaction parameters were entered into the full model: (f) f0 with each of the five measures of voice quality, (g) listener language with the five measures of voice quality, (h) speaker sex with the five measures of voice quality, (i) f0 with listener language and with speaker sex, and finally, (j) listener language with speaker sex. More complex interactions were not included in the model. A process of model comparison was then employed using log likelihood ratio tests in order to identify and remove non-contributing factors.[1]

This process resulted in a model containing 17 fixed-effects parameters listed in Table I, which shows the

TABLE I. Results of the statistical model of listeners' location-in-f0-range ratings. Δ LogLik indicates the decrement in model fit when a given parameter was removed from the chosen model, indicating its importance to the model. The five parameters most important to model fit are shown in bold.

Fixed effects	Estimate	Std. Error	t value	p value	Δ LogLik
(Intercept)	55.9899	3.5400	15.82	< 0.001	
f0	**0.7096**	**0.1260**	**5.63**	**< 0.001**	**−1781**
Sex (Male)	**−52.8520**	**5.3334**	**−9.91**	**< 0.001**	**−130**
CPP	**−1.7282**	**1.1902**	**−1.45**	**0.147**	**−84**
f0 × Sex (Male)	**0.6771**	**0.0524**	**12.91**	**< 0.001**	**−81**
F2	**0.8681**	**0.3253**	**2.67**	**0.008**	**−73**
CPP × Sex (Male)	21.1246	1.9545	10.81	< 0.001	−58
f0 × F2	−0.0943	0.0085	−11.06	< 0.001	−57
H1*-A3*	−3.4783	0.5855	−5.94	< 0.001	−49
f0 × H1*-A3*	0.1358	0.0158	8.60	< 0.001	−34
H1*-A3 × Sex	6.2027	0.7691	8.07	< 0.001	−33
F3	−0.1707	0.0839	−2.03	0.042	−23
F3 × Sex (Male)	0.8573	0.1409	6.08	< 0.001	−18
F2 × Sex	−2.2467	0.4538	−4.95	< 0.001	−13
f0 × CPP	0.2237	0.0490	4.57	< 0.001	−9
Listener Lang (Mand)	1.5143	2.0747	0.73	0.942	−5
Sex (Male) × Listener Lang (Mand)	1.4900	0.7691	1.94	0.053	−5
f0 × F3	0.0136	0.0033	4.08	< 0.001	−4
Random effects		Variance			Std. Dev.
Listener (Intercept)		41.0622			6.408
Speaker (Intercept)		5.6219			2.3711
Residual		272.6722			16.5128

parameters rank ordered according to their importance to the model (determined by the decrement in model fit when that parameter is removed from the model). We limit our discussion here to the five most influential parameters (shown in bold in the table): f0 (by far the most important parameter), speaker sex, CPP (Cepstral Peak Prominence), the interaction of f0 by speaker sex, and F2.

The effect of f0 was highly significant ($p < 0.001$), listeners' ratings of the stimuli increasing along with f0. The effect of speaker sex was also significant ($p < 0.001$), as was the interaction of f0 and speaker sex ($p < 0.001$), as shown in Fig. 3. In general, an f0 above approximately 250 Hz was associated with a higher location-in-f0-range rating when it came from the voice of a male speaker. There was also a significant effect for F2 ($p < 0.01$), and thus this formant frequency influenced listeners more strongly than the other formant frequencies did, though all contributed to the model. Figure 4 shows F2 as a function of f0, although this interaction was not one of the highly influential parameters in the model. Finally, CPP's effect was not significant here, but presumably contributed to model fit only via two low-ranked interaction terms (with speaker sex and with f0), and so we do not consider it further.

3. Judgments of synthetic stimuli

Listeners' responses were pooled for each of the synthetic tone stimuli as was done for the tokens above, and were plotted against the f0s of those synthetic tokens. The correlation of responses with stimulus f0 was quite high (R^2=0.961); as a group, listeners lined up the tone stimuli such that a tone stimulus with a higher f0 was rated as coming from a higher location in a hypothetical speaker's range, suggesting that listeners have expectations as to where specific f0s fall in such a range. The averaged responses are plotted in Fig. 5; no further modeling of these responses was carried out, as these synthetic tones possessed none of the other properties of voices of interest, and as there was so little variance remaining to be accounted for, beyond the effect of absolute f0.

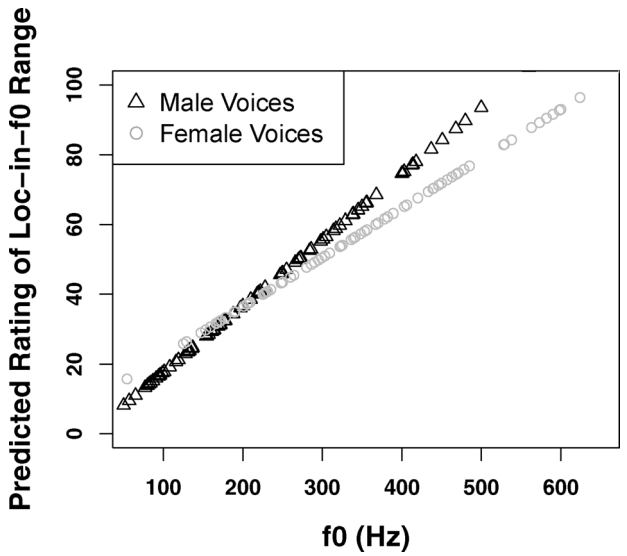

FIG. 3. Model predictions for f0 location ratings for male and female tokens, plotted as a function of f0; f0 values above the group mean (297 Hz) are associated with higher f0 location rating values when the voice is male rather than female.

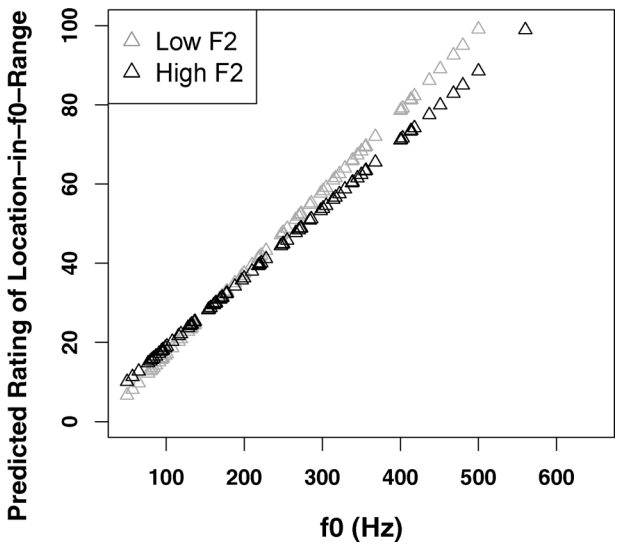

FIG. 4. Model predictions for f0 location ratings as a function of f0 at two different levels of the parameter "F2." The high value represents F2 values 1 standard deviation above the group mean, the "Low" value 1 standard deviation below the group mean.

C. Discussion

Honorof and Whalen (2005) had found that listeners' judgments regarding the location-in-f0-range of an unfamiliar voice sample were correlated with the actual location of that f0 in that speaker's range. Experiment 1 sought to replicate their basic result, and indeed listener ratings in the present experiment were generally well correlated with location-in-f0-range. We then sought to extend Honorof and Whalen's study by determining what information in the signal listeners rely upon. Listener judgments were most strongly related to absolute f0, a correlation which Honorof and Whalen had not tested. We then tested their suggestion that voice quality might also be a cue. However, our model, which tested a number of relevant measures, did not provide evidence for a major contribution by voice quality to listeners' judgments. This is not to say that voice quality played no role, as both Cepstral Peak Prominence and H1*-A3* contributed to the fit of the model. Nonetheless their contributions, indeed the contributions of all parameters except f0, speaker sex, and F2, were seen only in combinations with other parameters, and were not significant on their own. The model's results in this respect are not surprising given the correlations with listeners' ratings of the stimuli; the f0 of the tokens accounted for 88% of the variance in the ratings, leaving very little for other measures such as voice quality measures, and even speaker sex, to predict. While this might seem contrary to Honorof and Whalen's conclusions, our analyses, which provide novel pooled-listener correlations, correlations with absolute f0, and a statistical model that directly compared the contributions of individual acoustic cues, clearly indicate that when listeners make these judgments, they do so mostly on the basis of absolute f0.

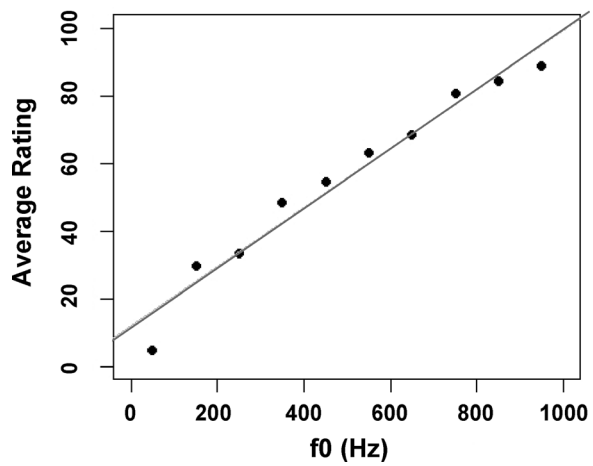

FIG. 5. Scatter plot showing averaged (pooled over listeners) "location-in-f0-range" ratings as a function of the f0 of each of the nine synthetic tone stimuli (R^2=0.96).

The main implication of this result is that when a listener makes a judgment about where an f0 should fall in the range of an unfamiliar speaker, that listener does not really place it in the range of that speaker; rather, an idealized speaker, or an entire population of speakers, is the basis for f0 normalization. That listeners have quite acute expectations about how certain f0s relate to speakers' ranges was also shown clearly by listeners' very orderly ratings of the non-voice tone stimuli. Our conclusion is that listeners' ratings of location-in-f0-range in our study, and therefore by extension also in Honorof and Whalen's (2005) study, had very little to do with the individual ranges of the speakers presented. Instead, experience-based knowledge of cross-speaker f0 ranges seems to be what matters.

Nonetheless, f0 is not used quite absolutely; instead there appear to be separate expectations for male voices and female voices, as discussed by Honorof and Whalen (2005, 2010) and Lee *et al.* (2010). In our model, the second most important factor after f0 was speaker sex, and its interaction with f0 was also a heavily weighted (relative to lesser factors) and highly significant factor. The model suggests that listeners treated f0s above approximately 250 Hz differently depending on the sex of the speaker, such that a given f0 was judged as coming from a higher location in the speaker's f0 range if it was produced by a male rather than a female. Again, this was evident in the correlation of responses with stimulus f0 in Fig. 2.

This interpretation implies that listeners were first able to make systematic decisions about the sex of speakers. It is extremely likely that they did this, albeit unconsciously: as reviewed by Kreiman and Sidtis (2011), sex identification of speakers from their voices is a basic and immediate biological response widespread among animals. However, we do not yet know how the listeners in experiment 1 might have accomplished this. "Speaker sex" was included in the model simply as a cover term for some (unknown) aspect(s) of the signal. As discussed in Sec. I A, there are various contributors to the perception of speaker sex, including f0 and vocal tract resonances, and potentially voice quality (e.g., Hanson and Chuang, 1999; Lee *et al.*, 2010; Honorof and Whalen, 2010; Kreiman and Sidtis, 2011). However, as Honorof and Whalen (2005, 2010) note, judgments of stimuli like those in experiment 1 are especially challenging because, unlike

stimuli used in most studies of sex perception, these stimuli were very brief, came from many speakers, and covered a wide range of f0s. Experiment 2 therefore investigated listeners' accuracy in identifying the sex of the speakers from the stimuli used in experiment 1, and also explored how the acoustic cues serve as predictors of their judgments.

III. EXPERIMENT 2

A. Methods

1. Stimuli

The stimuli were the same (voice only, not synthetic) as in experiment 1.

2. Listeners

Twenty-three native speakers of American English (mostly from California) and 23 speakers of Mandarin (either mainland or Taiwan) participated. None had participated in experiment 1, but all met the same selection criteria.

3. Procedure

The procedure for presenting stimuli and collecting responses for experiment 2 was very similar to that used in experiment 1. In experiment 2, however, the method of response required participants to click a button which appeared in the MATLAB GUI interface, rather than manipulate a slider bar. For each voice token presented, listeners were to select one of two buttons, one labeled "Male" and one "Female." Second, instruction was given to all subjects (native Mandarin and English-speaking) in English. Because all Mandarin-speaking participants were to some extent bilingual, this simple task was effectively explained by English-speaking experimenters. Other aspects of the presentation of the stimuli were the same as in experiment 1 (including blocking and randomization, no mention of the stimuli being from two languages, etc). Participants were given a practice session that confirmed that all understood what was being asked of them, and how to use the interface.

B. Results

1. Overall accuracy

To assess listeners' overall accuracy in identifying the sex of the speakers, the average probability of a correct response over each speaker's range of tokens was calculated. Collapsed over all speaker and listener groups, accuracy was on average 77.7% (SD=28). Accuracy values were also submitted to a three-way analysis of variance (ANOVA), with one between-subjects factor, listener language (English, Mandarin), and two within-subjects factors, speaker language (English, Mandarin) and speaker sex (male, female). A significant main effect was found for speaker language [$F(1,44)=50.15$, $p<0.0001$]; on average the probability of accurate responses was higher for English voices (80%, SD=11.4) than for Mandarin voices (75.4%, SD=13.1). There was also a significant main effect for speaker sex [$F(1,44)=14.48$, $p<0.001$]; on average female voices (81.9%, SD=10.1) were more likely to be correctly identified than male voices (73.4%, SD=13.2). Speaker language was also found to interact separately with listener language [$F(1,44)=27.01$, $p<0.0001$] and with speaker sex [$F(1,44)=62.77$, $p<0.001$]. Both English and Mandarin listeners identified the sex of voices best when those voices were English, but a Bonferroni pairwise comparison ($\alpha=0.05$) showed that while this difference was significant for English listeners (English voices: 82.3%, SD=8.8; Mandarin voices: 75%, SD=13), it was not for Mandarin listeners (English voices: 77.6%, SD=13.2; Mandarin voices: 75.9%, SD=13.4). Second, although there was no real difference in the identification of English male (80.8%, SD=12.2) vs English female (79.1%, SD=10.7) voices, there was a significant difference between Mandarin male (66%, SD=9.7), and Mandarin female (84.8%, SD=8.6) voices, and this difference held for both listener language groups.

These main effects and two-way interactions are best understood in terms of the significant three-way interaction between speaker sex, speaker language, and listener language [$F(1,44)=6.56$, $p<0.05$]. The interaction plot in Fig. 6 shows the pattern of sex identification accuracy for all speaker groups by both listener language groups. Numerically, there was a tendency for listeners to judge male voices more accurately when speaker language and listener language matched, this trend being strongest for English male voices (85%, SD=9.4 for English listeners compared with 76.5%, SD=13.3 for Mandarin listeners), and Bonferroni pairwise comparisons limited to the four speaker groups across the two listener groups show this to be the only significant difference. That is, the English listeners were more accurate on the English male voices than the Mandarin listeners were. Robustly significant, within both listener groups, however, was the disadvantage for Mandarin male voices compared with all other groups (64.6%, SD=8.6 for English listeners, 67.4%, SD=10.7 for Mandarin listeners).[4]

2. Correlations with f0

Whereas the ANOVA gave a picture of how listeners responded to the speakers' voices throughout their f0 ranges,

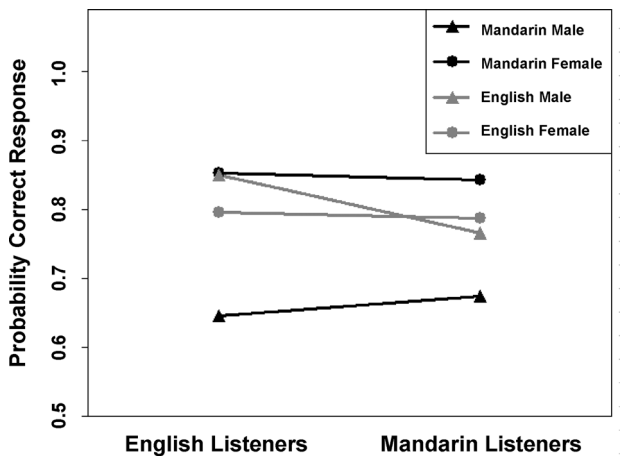

FIG. 6. Interaction plot showing the probability of accurate identification of speaker sex for the four speaker groups (pooled over the nine tokens of each speaker's range) by the two listener groups.

previous research shows that accuracy differs across f0s, being lower when a speaker is outside of the range typical of his or her sex. To examine how f0 might have affected accuracy here, correlations between f0 and listeners' accuracy were carried out for each of the four speaker groups, pooled across the two listener-language groups, shown in Fig. 7. As can be seen from the data plotted in this figure, there is a close relationship between f0 and accuracy for all four groups, broadly similar to the pattern found by Honorof and Whalen (2010). In general, female voices, especially Mandarin female voices, were most accurately identified as female at f0 values above 200 Hz, accounting for considerable portions of the variance in listeners' accuracy (R^2=0.38 for English female voices; R^2=0.43 for Mandarin female voices). Conversely, male voices were most accurately identified when f0 was below 200 Hz, accounting for a much larger portion of the variance in accuracy for these groups (R^2=0.84 for English males; R^2=0.73 for Mandarin males). This apparent f0-based response bias was shown to have implications for the statistical model of listeners' judgments of speaker sex, presented below.

3. Modeling listeners' judgments of speaker sex

As in experiment 1, we built a model of listeners' responses, in this case to determine which acoustic factors served as predictors to listeners' sex identifications. In particular, the outcome "probability of male response" was modeled using logistic regression, again including both fixed and random effects. Random effects included speaker and item; fixed effects tested in the model were similar to those used in experiment 1, and included the following 10 parameters: (a) the language of the listener (English or Mandarin); (b) f0; (c) resonances: the mean frequency of each of the first three formants; (d) mean measures of voice quality: CPP, H1*-A1*, H1*-A3*, H1*-H2*, H2*-H4*. In addition, 11 interaction parameters were tested: (e) the interactions of the five voice measures with f0 (f) listener language with f0, and (g) the five voice measures with listener language. The best-fitting model was chosen as in experiment 1, and included the following fixed-effects parameters: f0; F1, F2, and F3; H2*-H4*; H1*-H2*; H1*-A3*; listener language; the interactions of f0 with H2*-H4*, H1*-H2*, listener language, and with each F1, F2, and F3; the interaction of F1 and listener language. Table II shows the results of the model, with factors ranked by their importance to model fit as in experiment 1. By far, f0 had the greatest influence on model fit, followed by F2. The next most important parameters were H2*-H4*, its interaction with f0, and H1*-A3*. Although the individual contributions of each of the other factors in the model resulted in improved model fit, their influence is considerably smaller, and we again limit our discussion to these five most important parameters, shown in bold in the table.

The results of the model show a significant main effect of f0 on the probability of "male" responses ($p < 0.001$),

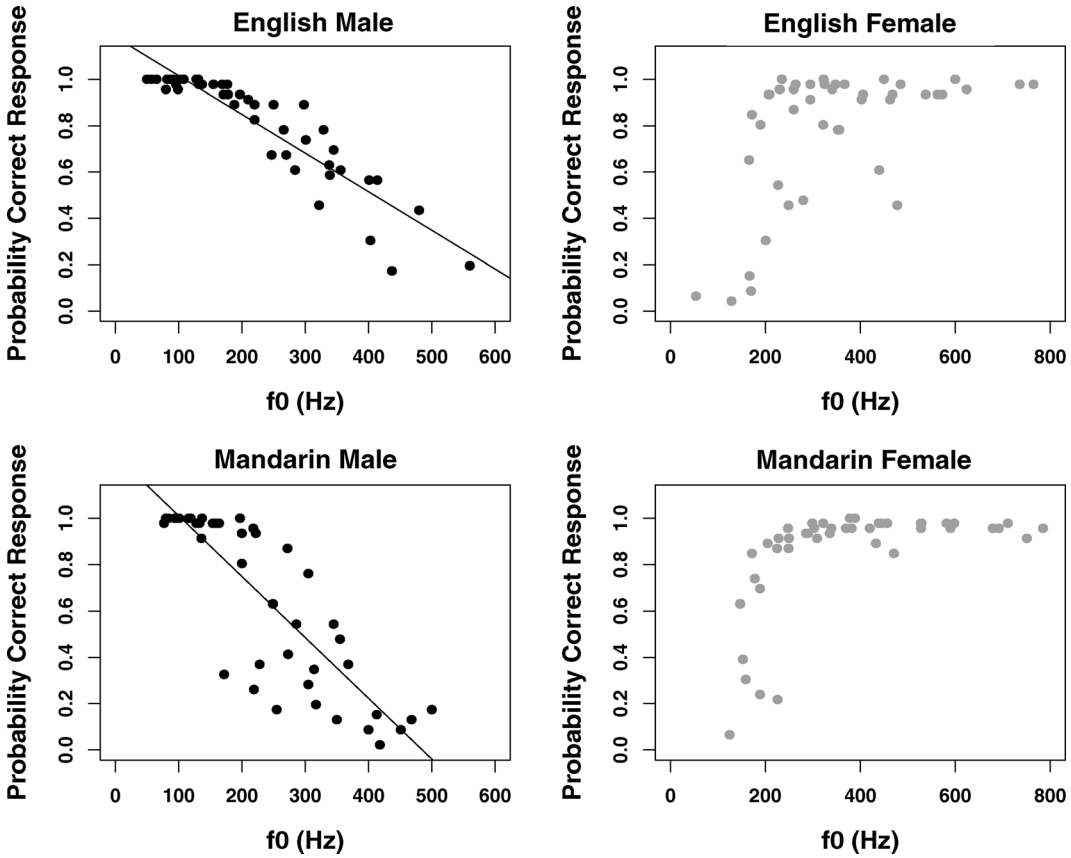

FIG. 7. Probability of correct identification of speaker sex as a function of f0 for the four speaker groups: English males (R^2=0.84), Mandarin males (R^2=0.73), English Females (R^2=0.38), and Mandarin females (R^2= 0.43). Regression lines are not shown for the female data because the fits are relatively poor.

TABLE II. Results of the statistical model of "male" responses in the sex identification experiment. Δ LogLik indicates the decrement in model fit when a given parameter was removed from the chosen model, indicating its importance to the model. The five parameters most important to model fit are shown in bold.

Fixed effects	Estimate	Std. Error	t value	p value	Δ LogLik
(Intercept)	−0.1955	0.6542	−0.299	0.765	
f0	**−0.3276**	**0.0305**	**−10.727**	**< 0.001**	**−368**
F2	**0.0511**	**0.0425**	**1.203**	**0.229**	**−30**
H2*-H4*	**0.0627**	**0.0802**	**0.781**	**0.435**	**−12**
f0 × H2*-H4*	**−0.0263**	**0.0054**	**−4.829**	**< 0.001**	**−12**
H1*-A3*	**−0.4884**	**0.1154**	**−4.233**	**< 0.001**	**−10**
H1*-H2*	0.1437	0.1019	1.410	0.159	−6
f0 × H1*-H2*	−0.0271	0.0077	−3.500	< 0.001	−6
Listener Language (Eng)	0.4671	0.3877	1.205	0.228	−5
F1	−0.1407	0.0587	−2.401	0.016	−5
f0 × Listener Language (Eng)	0.0165	0.0061	2.727	0.006	−4
F3	−0.0060	0.0173	−0.371	0.711	−4
f0 × F2	0.0185	0.0024	7.734	< 0.001	−3
f0 × F3	−0.0031	0.0013	−2.485	0.013	−3
F1 × Listener Language (Eng)	−0.0464	0.0427	−1.086	0.277	−1
f0 × H1*-A3*	0.0061	0.0066	0.924	0.355	−1
Random effects		Variance			Std. Dev.
Listener (Intercept)		0.6222			0.7888
Speaker (Intercept)		3.1193			1.7662

such that higher f0s were strongly associated with a lower probability of male responses. Although F2 and H2*-H4* were both important to model fit, neither showed significant main effects (both $p > 0.2$). However, the interaction of f0 and H2*-H4* was significant ($p < 0.001$), the largest effect being that at f0s below approximately the group mean (297 Hz), higher values of H2*-H4* were associated with a higher probability of a "male" response. F2 also entered into an interaction with f0 that indicated F2 to be inversely related to "male" responses, but only at higher f0s. While this interaction was much less important to the model, it presumably accounted for F2's ranking in the model. The fifth and final parameter considered here was H1*-A3*, for which there was a significant main effect ($p < 0.001$). When other parameters in the model are held at their mean values, lower values of H1*-A3* were associated with a higher probability of "male" responses.

4. Differences between groups

One of the findings in experiment 2 was that the listeners' identification of speaker sex was not equally accurate for all groups of speakers. Our findings about the aspects of the signal that predicted listeners' decisions allow us to characterize the "difficult" voices, that is, the voices whose sex was less well identified. We focus our discussion here on the least-well identified group of speakers, Mandarin males, who were remarkably-poorly identified by both English and Mandarin listener groups, hardly much above chance.

The first and most obvious explanation for why the Mandarin males might not have been easily identified as male would be because they were not, at least in our stimuli, prototypical in terms of what listeners primarily based their responses on, namely f0 (see, e.g., Honorof and Whalen, 2010). This turns out to be consistent with the properties of these stimuli. Figure 8 shows the f0 for each token taken from the ranges of the five Mandarin male speakers, ordered from least-well identified (Speaker 14) to best identified (Speaker 13), compared to the best-identified English male speaker (Speaker 4). As the figure shows, compared to other male speakers, the Mandarin male speakers who were hardest for listeners had higher f0s for each of the tokens in their range, save the lowest token. Indeed, except for Speaker 12, f0 is an almost perfect predictor of how difficult a given Mandarin male speaker was to identify relative to other Mandarin speakers. Again, this is in agreement with the general theme of both experiments presented above, which is listener attentiveness primarily to f0.

However, the relationship between f0 and listeners' judgments of the Mandarin male voices was not perfect. We therefore explored the Mandarin male stimuli further, to examine what acoustic properties other than f0 distinguished the most difficult from easier tokens. A subset of the Mandarin male tokens was selected—those with f0s between 150 and 350 Hz—and within that f0 range further divided into two groups on the basis of the regression line seen in Fig. 7 (bottom left, Mandarin male data), which reflects the fairly linear relation for this subset of tokens. Tokens above the line were considered well-identified, while tokens below the line were considered poorly identified. The mean values for the acoustic parameters in the chosen model were calculated for just these groups, and compared between the two groups. Note that f0's explanatory role should be relatively reduced in this comparison, since the tokens being compared came from within the same limited range of f0s. These two groups of tokens differed considerably on three acoustic measures. Both F2 and F3 tended to be higher for the poorly-identified tokens, compared with the well-identified tokens (1454 Hz

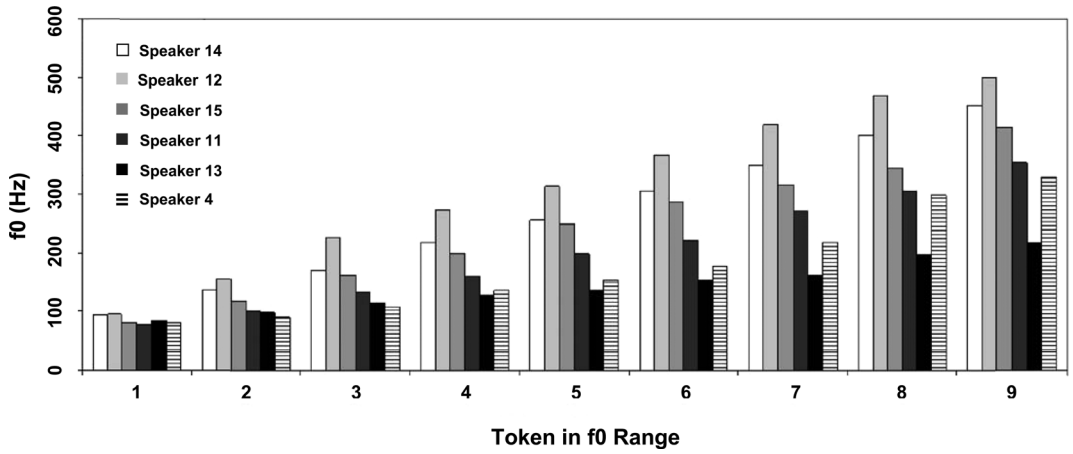

FIG. 8. F0s for each of the tokens from the five Mandarin male speakers, ordered from the least-well identified speaker (Speaker 14) to the best-identified speaker (Speaker 13). Also shown is the best-identified English male speaker (Speaker 4). F0 is a good predictor of accurate sex identification for most tokens for most Mandarin male speakers.

versus 1213 Hz for F2, 2959 versus 2706 for F3), and the same pattern held for one of the voice measures, H1*-A3* (5.24 dB versus −2.4 dB). This provides further evidence that listeners were attentive to these aspects of the signal, since for these tokens, they led listeners to the wrong decisions.

Note that the difficulty with the Mandarin male voices was equal for English and Mandarin listeners (see Fig. 6 and related discussion). Thus this difficulty cannot be attributed to differences in the experience of the two listener groups with Mandarin male voices, or with any particular voice properties that the larger population of Mandarin males may employ, either typically or occasionally. The effect must arise from the properties of these particular stimuli: for whatever reason, four of our five Mandarin male speakers are apparently not typical.

Nonetheless, there was one effect of listener language in our data: English listeners did significantly better with English voices (compared to Mandarin) while Mandarin listeners performed similarly on the two languages. The English listeners had significant prior experience only with English voices, while the (somewhat bilingual) Mandarin listeners had prior experience with both languages. If prior exposure is the basis for expectations about a population, then the English listeners' expectations about male voices would be optimized to English, and not as good a fit to the Mandarin voices. In contrast, the Mandarin listeners' expectations about male voices derive from combined speaker populations and are likely a compromise between the two languages.

C. Discussion

The purpose of experiment 2 was to determine two aspects of the stimuli used in experiment 1: how well the listeners in experiment 1 could have identified the sex of the speakers, thus replicating Honorof and Whalen (2010) and Lee *et al*. (2010), and on what signal properties the listeners in experiment 2 based their decisions. With respect to accuracy, overall accuracy of identification here was about 78%, slightly higher than Honorof and Whalen's (2010) accuracy of about 72%, but as in their study, accuracy varied with stimulus f0. (Lee *et al*.'s (2010) listeners were much more accurate, at about 90%, but for the more limited f0 range of natural speech.) That is, listeners can certainly judge speaker sex well above chance, and thus could have done so in our experiment 1; but they do make mistakes, and thus in experiment 1 they would have sometimes used the wrong f0 population range and misplaced some tokens in their f0 ranges. Like Honorof and Whalen and Lee *et al*. we found that when f0 is high, male voices are mistaken for female, and in our experiment 1 male stimuli are never rated as being at the top of their ranges.

With respect to the signal properties used, consistent with many previous studies (see references in Kreiman and Sidtis, 2011), f0 was the most important predictor of our listeners' judgments. This is particularly true for the English male voices, which showed a quite linearly declining level of accuracy as f0 increased. In contrast, for the female voices there was a less linear relationship between f0 and accuracy. This difference is understandable for the stimuli used in the experiment, given the distributions of the f0s of the two sexes. As can be seen in Fig. 1, f0s below about 175 Hz are almost all male, while those above 500 Hz are almost all female. Only about half (43/90) of male f0s are above 200 Hz, while 75/90 female f0s are. F0s of about 200–350 Hz are ambiguous, but relatively high for the men. Figure 7 indicates that listeners used 200 Hz as a fairly strict threshold for female f0, but a more gradual criterion for male f0 in this ambiguous region.

The second most important factor to our model of listeners' judgments of sex was the frequency of F2, also found to be an important predictor in previous studies. It was also apparent upon closer inspection of the subset of poorly identified Mandarin male tokens, that these tokens were characterized by higher F2 and F3 values. The importance of F2, as opposed to F3 or any other formant, as a predictor possibly means that in this particular sample of voices, F2 happened to be more saliently different across the speakers, so that listeners gave it more weight.

There was also an effect of some of the voice quality measures tested in the model. Honorof and Whalen (2010) found no correlation between their voice quality measures

and listeners' judgments, but they tested only jitter and shimmer measures. Previous production studies have found female voices to be associated with higher values on measures of breathiness and spectral tilt (e.g., Klatt and Klatt, 1990; Hanson, 1997; Hanson and Chuang, 1999; Lee, 2009), and our results provide evidence that in fact spectral tilt (in the form of the measure H1*-A3*) represents an aspect of the signal listeners attend to when making judgments regarding a speaker's sex. Higher values for H1*-A3* were associated with female rather than male responses in the model and, again, inspection of a subset of tokens suggested that Mandarin male speakers with higher values of H1*-A3* were difficult to identify as male. Finally, H2*-H4* was the most important factor in our model after f0 and F2, and was highly significant. At least for a subset of f0s, higher H2*-H4* values were associated with a greater likelihood of a "male" response by listeners. This direction of effect is unexpected, the opposite of the pattern seen in the stimuli themselves, where H2*-H4* patterns like H1*-A3*. It seems that in the model, H1*-A3* accounts for variance shared by these two factors, and the H2*-H4* $\times$ f0 interaction fine-tunes the model beyond what H1*-A3* can account for. Possibly in the f0 range around 150 Hz, higher H2*-H4* reflects not breathiness, but the absence of the somewhat creaky voice that females would typically produce at these lower f0s.

IV. GENERAL DISCUSSION

This study sought to provide some evidence regarding how listeners in Honorof and Whalen's (2005) perception experiment could have managed to place an individual f0 within a speaker's range—with no prior experience with that range, no syllable-external information, and no dynamic syllable-internal f0 information on which to base f0 normalization. The apparent implication of their result was that the listeners used other signal-intrinsic information in addition to f0 to make their decisions. One hypothesis of special interest was that voice quality could be such a source of information. In addition, we also wished to distinguish three different ways that normalization could utilize voice quality cues. The first is directly for f0 range, as when voice quality differs between each speaker's low vs high pitches, consistently across speakers. In this case voice quality would be an excellent cue for normalization of f0 range. The second way is directly for absolute f0, as when voice quality co-varies with f0. In this case voice quality would be a redundant cue for f0 and thus provide no additional information about f0 range beyond f0 itself. The third way is indirectly, as when voice quality cues some other property, such as speaker sex, that could be used in normalization.

The results of our experiments do not seem to provide support for either of the direct uses of voice quality. First, the statistical model of listeners' ratings from experiment 1 indicated that voice quality, or other signal-intrinsic information beyond f0 itself, is hardly needed to perform this task. Absolute f0 was by far the most important predictor of listeners' ratings, and was much better correlated with mean ratings than was the actual location-in-f0-range (the presumed "right answer"). We believe that the (weaker) correlation with location-in-f0-range is simply a consequence of the fact that location-in-f0-range is itself necessarily somewhat correlated with absolute f0, and therefore will also be somewhat correlated with the ratings.

Indirect use of voice quality information to judge location-in-f0-range was, however, robustly evident; listeners' decisions about the location of an f0 in a speaker's range were partially dependent on the sex of the speaker. The results of experiment 1 suggested that listeners used two population distributions of f0, male and female, when interpreting the stimuli. The results of experiment 2 then suggested that listeners' decisions about the sex of the speaker were dependent on a number of acoustic parameters, including voice quality, though again f0 was the most relevant. The voice quality measures found to be most relevant for sex identification were H1*-A3* and H2*-H4* (especially, its interaction with f0). Thus f0 was used both directly (listeners know the location-in-range of a given f0) and indirectly (by providing a basis for sex identification). In this way our study provides an explicit statistical model that accords with some aspects of the suggestions by Lee (2009) and Lee *et al.* (2010) about Honorof and Whalen's (2005) findings.

That listeners associated higher values of H1*-A3* (which perhaps index reduced speed or abruptness of glottal closure) with female rather than male voices would be predicted from previous studies. H2*-H4*, however, is less clearly understood. Our results suggest that H2*-H4* plays a more prominent role in listeners' perception of speaker sex than any other, more common, measure of voice quality, but in possibly unexpected ways. Characterizing the aspect(s) of the voice reflected by H2*-H4* is thus a necessary task for future research.

ACKNOWLEDGMENTS

This research was supported by NSF grant BCS-0720304 to the second author. A preliminary version of this work was presented at the Spring 2010 meeting of the Acoustical Society of America in Baltimore. We thank Yen-Liang Shue for all his help with VoiceSauce; UCLA undergraduates Grace Tsai and Niloofar Yaghmai for help in carrying out the experiments; Bruce Gerratt for suggesting the tone condition in experiment I; Jody Kreiman and Abeer Alwan for helpful discussions; and the statistics consultants of the UCLA Academic Technology Service.

[1]See supplementary material at http://dx.doi.org/10.1121/1.4714351 for additional description of stimulus generation and statistical modeling.

[2]It is also important to note that, because the frequencies of H2 and H4 are twice and four times the fundamental frequency, then (assuming a linear source-filter model) H2-H4 is especially sensitive to the influence of the formant frequencies; as a result, corrected H2*-H4* values can be quite different from uncorrected H2-H4. "Source H2-H4" here refers to acoustic measures of the source function, where no correction is needed.

[3]We also explored the possibility that, during the course of the experiment, listeners learned something about the voices they were hearing. We compared the individual (not group) R^2s for tokens that were the first or last exposure to a speaker's voice. In fact, average R^2 values were relatively high for first-exposure tokens (on average R^2=0.391, SD=0.169), indicating that listeners were able to perform the task without experience from

previous tokens of speakers' voices. Still, the relationship was somewhat stronger for last-exposure tokens (average R^2=0.445, SD=0.187). Whether this change involved learning about individual voices, about the set of voices as a whole, or simply about the task itself, cannot be known. However, since the difference was only marginally significant (paired t-test: t(40)=−1.56, p=0.06), it is unlikely to account for any of the primary findings.

[4]As in experiment 1, we compared performance on first versus last exposures to the voices. On average, accuracy was relatively high for both: 78.5% (SD=8%) for first-exposure and 78.3% (SD=7.9%) for last-exposure tokens. This difference was not significant (paired t-test, [t(45)=0.131, $p > 0.1$]), indicating that repeated exposure to the voices throughout the course of the experiment did not benefit listeners' identification of speaker sex.

Baayen, R. H. (**2008**). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R* (Cambridge University Press, Cambridge), pp. 253–259.

Baken, R. J., and Orlikoff, R. F. (**2000**). *Clinical Measurement of Speech and Voice* (Singular, San Diego), pp. 185–187.

Blankenship, B. (**2002**). "The timing of nonmodal phonation in vowels," J. Phonetics **30**, 163–191.

Boersma, P., and Weenink, D. (**2008**). "Praat: Doing phonetics by computer (version 5.0.13) [computer program]," http://www.praat.org (Last viewed June 18, 2012).

Esposito, C. M. (**2010**). "The effects of linguistic experience on the perception of phonation," J. Phonetics **38**, 306–316.

Greenberg, S., and Zee, E. (**1979**). "On the perception of contour tones," UCLA Working Pap. Phonetics **45**, 150–164.

Hanson, H. M. (**1997**). "Glottal characteristics of female speakers: Acoustic correlates," J. Acoust. Soc. Am. **101**, 466–481.

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.

Henton, C. G., and Bladon, R. A. (**1985**). "Breathiness in normal female speech: Inefficiency versus desirability," Lang. Commun. **5**, 221–227.

Hillenbrand, J., Cleveland, R. A., and Erickson, R. L. (**1994**). "Acoustic correlates of breathy vocal quality," J. Speech Hear. Res. **37**, 769–778.

Hillenbrand, J., and Houde, R. A. (**1996**). "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," J. Speech Hear. Res. **39**, 311–321.

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.

Honorof, D. N., and Whalen, D. H. (**2010**). "Identification of speaker sex from one vowel across a range of fundamental frequencies," J. Acoust. Soc. Am. **128**, 3095–3104.

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**, 2283–2295.

Keating, P., Esposito, C., Garellek, M., Khan, S., and Kuang, J. (**2011**). "Phonation contrasts across languages," in *Proceedings of the International Congress of Phonetic Sciences XVII*, edited by W.-S. Lee and E. Zee (City University of Hong Kong, Hong Kong), pp. 1046–1049.

Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857.

Kreiman, J., Gerratt, B. R., and Antoñanzas-Barroso, N. (**2007**). "Measures of the glottal source spectrum," J. Speech Lang. Hear. Res. **50**, 595–610.

Kreiman, J., and Garellek, M. (**2011**). "Perceptual importance of the voice source spectrum from H2 to 2 kHz," J. Acoust. Soc. Am. **130**, 2570.

Kreiman, J., and Sidtis, D. V. L. (**2011**). *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception* (Wiley-Blackwell, Hoboken, NJ), Chap. 4, Sec. 4.3.

Kuang, J. (**2011**). "Phonation contrast in two register contrast languages and its influence on vowel and tone," in *Proceedings of the International Congress of Phonetic Sciences XVII*, edited by W.-S. Lee and E. Zee (City University of Hong Kong, Hong Kong, China), pp. 1146–1149.

Laver, J. (**1980**). *The Phonetic Description of Voice Quality* (Cambridge University Press, Cambridge), Audio cassette illustrations for Chap. 3.

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

Lee, C.-Y. (**2009**). "Identifying isolated, multispeaker Mandarin tones from brief acoustic input: A perceptual and acoustic study," J. Acoust. Soc. Am. **125**, 1125–1137.

Lee, C.-Y., Dutton, L., and Ram, G. (**2010**). "The role of speaker gender identification in relative fundamental frequency height estimation from multispeaker, brief speech segments," J. Acoust. Soc. Am. **128**, 384–388.

Mecke, A.-C., Sundberg, J. (**2010**). "Gender differences in children's singing voices: Acoustic analyses and results of a listening test," J. Acoust Soc. Am. **127**, 3223–3231.

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

Perkell, J. S., Hillman, R. E., and Holmberg, E. B. (**1994**). "Group differences in measures of voice production and revised values of maximum airflow declination rate," J. Acoust. Soc. Am. **96**, 695–698.

Podesva, R. J. (**2007**). "Phonation type as a stylistic variable: The use of falsetto in constructing a persona," J. Sociolinguistics **11**, 478–504.

Reich, A. R., Frederickson, R. R., Mason, J. A., and Schlauch, R. S. (**1990**). "Methodological variables affecting phonational frequency range in adults," J. Speech Hear. Dis. **55**, 124–131.

Shue, Y.-L. (**2010**). *The Voice Source in Speech Production: Data, Analysis and Models*, doctoral dissertation (University of California, Los Angeles, CA), Chap. 4, Sec. 4.4.

Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. (**2011**). "VoiceSauce: A program for voice analysis," in *Proceedings of the International Congress of Phonetic Sciences XVII*, edited by W.-S. Lee and E. Zee (City University of Hong Kong, Hong Kong), pp. 1846–1849.

Stevens, K. N., and Hanson, H. M. (**1995**). "Classification of glottal vibration from acoustic measurements," in *Vocal Fold Physiology: Voice Quality Control*, edited by O. Fujimura and M. Hirano (Singular, San Diego), pp. 147–170.

Swerts, M., and Veldhuis, R. (**2001**). "The effect of speech melody on voice quality," Speech Commun. **33**, 297–303.

Various contributors (**2008**). "Audacity (Version 1.3.4-beta) [computer program]," http://audacity.sourceforge.net (Last viewed May 2008).

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.

Zhang, Z., Kreiman, J., and Gerratt, B. R. (**2011**). "Perceptual sensitivity to changes in vocal fold geometry and stiffness," J. Acoust. Soc. Am. **129**, 2529.

Zraick, R. I., Nelson, J. L., Montague, J. C., and Monoson, P. K. (**2000**). "The effect of task on determination of maximum phonational frequency range," J. Voice **14**, 154–160.