UCLA Working Papers in Phonetics, No. 108, pp. 113-140

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

Jason Bishop and Patricia Keating

(j.bishop@ucla.edu, keating@humnet.ucla.edu)

ABSTRACT

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 samples 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 determinant of listeners' judgments, but now voice quality measures also played a role. Thus it seems that pitch level judgments depend on voice quality only indirectly, through its information about sex. Absolute f0 is the most important information for deciding both pitch level and speaker sex.

I. INTRODUCTION

Fundamental frequency (f0) conveys linguistic information intended by the speaker, but it does so simultaneously with paralinguistic (e.g., emotion, emphasis) and non-linguistic (physiology) information about the speaker, and so the listener's interpretation of the linguistic message depends crucially on her ability to identify a speaker's purposeful highs and lows in f0. This ability is obviously relevant in tone languages, in which lexical specifications are partly encoded in f0. Its importance is not limited to tone languages, however, as highs and lows in f0 are used by both tone and non-tone languages to express intonational meaning. The main question we ask here is *how* listeners are able to disentangle a speaker's purposeful use of f0 from the non-linguistic, speaker-dependent way it is transmitted.

Other recent research has examined exactly this issue. Honorof and Whalen (2005), for example, examine—and rule out—some of the most intuitive possibilities, the first of which is that listeners rely on familiarity with an individual speaker's voice. The idea here is that exposure to various f0s produced by a speaker allows for estimation of that speaker's overall f0 range, which in turn allows a listener to place a given f0 accordingly in that range. Indeed, when experience (including sentential context) is available, listeners are known to use it for this purpose (Leather, 1983; Wong and Diehl, 2003; Moore and Jongman, 1997). However, Honorof and Whalen presented listeners with very brief (500ms), isolated, steady-state voice samples,

taken from various locations throughout the f0 ranges of twenty different (unfamiliar) English speakers. When presented with these brief samples, English-speaking listeners were able to place them in the ranges of the individual speaker's they came from with some accuracy —'accuracy' gauged by a correlation between listeners' judgments of location in f0 range of the tokens and their actual locations in the speakers' ranges (as determined in a separate production task). Thus, it cannot be the case that listeners require experience with an individual speaker's voice in order to know her overall range, and use that information in turn for identifying what is 'high' and what is 'low' for that speaker. Importantly, because they presented level, steady-state tones in isolation to listeners, Honorof and Whalen rule out a second possible basis for normalization that listeners might well make use of: cues from dynamic information about f0. That is, it is conceivable that listeners could estimate a larger range for a speaker based on a contour through a subsection of that range (for relevant results, see Moore and Jongman, 1997).

The implication of Honorof and Whalen's results, then, is that listeners may have access to signal-internal cues other than f0 that are informative as to a speaker's f0 range. As to what this information might be, the authors note that voice quality could conceivably serve as a cue, citing evidence from production studies (e.g., Swerts and Veldhuis, 2001) that certain voice quality characteristics may systematically co-vary with f0. Although they do not test the hypothesis in their study, Honorof and Whalen discuss the possibility that voice quality provided listeners with information regarding the f0 ranges of individual voices used in their study, allowing listeners to place a given f0 accordingly. A second possibility these authors note, however, is that listeners might have been able to identify the sex of the speakers in their experiment, allowing them to use this information to make sex-specific decisions as to speaker's f0 locations, relying on experience-based knowledge with overall f0 ranges typical of each sex. Honorof and Whalen also make clear that the use of signal-intrinsic correlates of f0 range (related to voice quality or not) and the use of experience-based templates of a population f0 range need not be mutually exclusive; both may be used simultaneously for judgments about location of an f0 in a speaker's individual range, and indeed if identification of the sex of a speaker is involved, this itself is a signal-intrinsic cue to f0 range, albeit a more indirect one. That is, listeners' judgments of f0 location would depend on their identification of the speakers' sex, which would need to be based on properties of the signal.

Results relevant to Honorof and Whalen's are reported by Lee (2009), whose study involved a more linguistic task. Lee presented forty native Mandarin-speaking listeners with the syllable /sa/, produced with each of the four Mandarin tones by 32 (16 male and 16 female) Mandarin speakers. These /sa/ syllables were edited such that listeners heard only the fricative and the first six glottal pulses of the vowel. This manipulation effectively neutralizes dynamic f0 information about the tones (Greenberg and Zee, 1979), but leaves the onset f0 information intact. If listeners were able to perceive this very brief f0 information about the high or low onset of the tone, it might allow them to distinguish between /sa/ produced with either Tone 1 (high level) or Tone 4 (high falling) from /sa/ produced with either Tone 2 (mid rising) or Tone 3 (low rising). In fact, Lee's listeners showed an ability to distinguish the high from low onset tones, which would only be possible with knowledge of what is 'high' and what is 'low' for the relevant (and also unfamiliar) speaker. While this result, like Honorof and Whalen's, provides further evidence that listeners were able to identify speakers' individual ranges, it does not answer the question of how they do this.

Lee (2009) suggests that both voice quality and gender detection could have been implicated. F0 and three measures associated with voice quality, H1-A1, H1-A3, and H1-H2,

were included in regression analyses of (a) listeners' accuracy in identifying individual tones, and (b) identification of 'high' versus 'low' onset tones. For individual tones, these four factors accounted for only 6.1% of the variance, no factor being significant. In the case of the high/low onset distinction, 21.1% of the variance was accounted for, and f0 was the best and only significant factor. While the author suggests that co-variation between the voice quality measures and f0 were used by listeners, the regression analyses presented did not seem to support such a strategy. A further suggestion, however, was that sex identification may have provided information as to whether tones should be categorized by listeners as low or high: listeners first identify the sex of the speaker, and then apply experience-based knowledge of sex-specific f0 ranges (that is, population ranges stored in memory). Limiting the hypothesized f0 range by sex in this way would in turn allow for an improved ability to identify high versus low. Note that, although the use of experience does not exclude the ability to extract information about f0 range from the signal, to the extent that the listener is able to use experience-based knowledge, attention to signal information becomes somewhat less important (and vice versa).

A. Voice quality as a cue to location in f0 range

The studies just discussed provide evidence that listeners do not need familiarity, context, or dynamic information about f0 to make judgments about f0 location in range. In this paper we try to identify the acoustic properties of voices relevant to performing this task. A particular hypothesis that we are interested in exploring is that voice quality is one property useful to listeners. However, we wish to make explicit what we understand to be the possible ways listeners could make use of voice quality, or any other aspect of the acoustic signal, for the purpose of determining a speaker's individual f0 range. The first is a *direct* method. If voice quality is useful for recovering the location of an f0 in an individual speaker's own range, it means there is a sufficiently salient relationship between a value on acoustic parameter X and a speaker's location in her own individual f0 range, such that value Y on acoustic parameter X indicates location Z in range. From the listener's perspective this would in effect mean (taking H1-H2 as an example), "this H1-H2 value is low, therefore this f0 is low in the speaker's range". Frequent comments found in the literature make reference to a co-variance between various measures of voice quality and f0. Extremely important to emphasize here is that for voice quality to be useful in this direct sort of manner, the observed correlation must hold *within* speakers (and consistently so from speaker to speaker). This is the kind of correlation presented by Swerts and Veldhuis (2001) for f0 vs. H1-H2, but these within-speaker correlations were very variable, indeed in opposing directions, across the speakers in their study. In contrast, correlations *across* speakers between f0 and H1*-H2* were demonstrated (for men) by Iseli et al. (2007), but this kind of correlation cannot help in locating an f0 in an individual speaker's range. To see this, consider a scenario in which voice quality measure X is perfectly correlated with f0. In such a case, any given value Y on voice measure X would tell a listener nothing at all about location Z in the speaker's f0 range that f0 would not; assuming that f0 is sufficiently salient, voice measure X would be entirely redundant and therefore uninformative. Taking Iseli et al.'s result as an example, a low value of H1*-H2* merely tells you that that sample has a low f0 for a man, not that the f0 is low for that individual man. Thus, a prerequisite for a direct use of any parameter of voice quality is that it varies along a speaker's f0 range, and that it does so more reliably than it does with f0 across speakers.

The second way voice quality could be used is in a more *indirect* manner, one in which voice quality serves as a cue to some other aspect of the speaker that allows the listener to assume an f0 range already stored in memory. Such a "known" range might be that of an individual familiar speaker (i.e., in speaker identification) or might be that of a group, based on experience with (and generalizations regarding) a population. As an illustration of how this use of voice quality could be implemented, consider a very relevant group: 'male speakers'. Presented with the relevant acoustic information, the listener might assume "this H1-H2 is low, so this must be a male voice; since this f0 is high for the average male, it must be high for this speaker". Obviously, such a strategy depends on the listeners' ability to assign the speaker to the proper group so that the appropriate known range can be referenced. In the present case, this implies identification of speaker sex. The relevant question then becomes "is voice quality sufficiently useful for determining speaker sex?".

While we do not know of a study that has directly tested listeners' use of voice quality in sex identification, there are a number of studies that indicate the signal contains differences that could be exploited. Henton and Bladon (1985) and Klatt and Klatt (1990), for example, show that H1-H2 values are consistently higher in female voices than male voices, and Perkell et al. (1994) show that this is also the case for spectral tilt (as reflected in the measure H1-A3). In their comparison of male and female voices, Hanson and Chuang (1999) replicate each of these findings, and additionally show that females show higher values for H1-A1. Shue (2010) found that voice measures can improve automatic gender classification, though significantly so only for 10-14 year old children's voices. It is thus plausible that the aspects of the voice that these measures reflect contributed meaningfully to listeners' ability to make decisions regarding f0 range in Honorof and Whalen's (2005) and Lee's (2009) experiments—not directly by way of indicating the location in a given speaker's own range per se, but more indirectly by helping to identify the sex of the speakers, and thus the sex-based f0 distributions to which a given f0 token belonged. Indeed, this was in part the suggestion made in Lee et al. (2010). In his follow-up sex-identification experiment, Lee et al. showed that both Mandarin and English-speaking listeners were able to identify the speaker sex of the tokens used in Lee (2009). Since the male and female tokens were somewhat distinguished by some measures of voice quality, Lee et al. reasoned that voice quality was in fact used for sex identification. While this conclusion is consistent with the findings, it is not the necessary conclusion, as listeners have other possible—and perhaps more salient—cues to a speaker's sex that could be equally appealed to in explaining the results. 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 either of these two aspects of the signal are above those found in the roughly normal speaking range of the average male. In fact, Lee et al. (2010) does find an asymmetry in the accuracy results in his study that seems indicative of such a bias: sex identification of female speakers was better when the onset tone stimulus was high, and the opposite pattern held for male voices. This suggests that listeners were making decisions about speaker sex based in part on f0. To establish that voice quality was contributing to listeners identification of speaker sex (and thus indirectly to the speaker's f0 range) it is necessary to show that voice quality independently accounts for some portion of the variance in listeners' judgments. The problem is essentially analogous to the one of assuming that listeners exploit a correlation between certain voice quality characteristics and f0 to identify f0 location; for voice quality to be useful for identifying speaker sex, it must account for some of the variance that is

not accounted for by the other (possibly more auditorily salient) parameters that also correlate with speaker sex.

B. Present study

In the two experiments below, we explore what factors contribute to a listener's placement of an f0 in the ranges of individual Mandarin and English speakers, using the kind of brief stimuli presented to listeners in Honorof and Whalen's (2005) study. In so doing, we are particularly interested in Honorof and Whalen's hypothesis that voice quality is implicated in listeners' performance, either indirectly or directly (or both), as discussed above. In Experiment 1, we attempt to replicate the basic findings reported in Honorof and Whalen (2005) regarding listeners' ability to locate f0s in speaker-specific ranges. We do this with the goal of building a model of the listeners' performance based on a range of possible cues in the signal, a number of which we take to reflect voice quality. The conclusion we draw on the basis of that model is that listeners' performance on the f0-location identification task can be explained largely by f0. However, there is a highly significant interaction between f0 and the sex of the speaker, indicating that listeners' use of f0 is modulated by their knowledge of the f0 range of the average speaker of each sex, and implying that something in the signal indicates the sex of the speaker. In Experiment 2 we examine what factors distinguish the sex of the speakers, and we investigate this by way of a model as well, based on a set of parameters similar to that in Experiment 1. The conclusion we draw from that model is that voice quality is in fact useful for identifying speaker sex, although f0 itself is again by far the most important factor.

Taken together, the data suggest that the role voice quality plays in Honorof and Whalen's (2005) findings is best characterized as indirect. At least when the listener lacks familiarity and is denied context, voice quality is one of the cues that listeners can use to make decisions about a speaker's sex; knowledge of the speaker's sex in turn allows for a more specific decision about where a given f0 should fall in the range of the speaker.

II. EXPERIMENT 1

A. Method

1. Participants

Ten adult native speakers of English and 10 of Mandarin (5 males and 5 females of each language) participated in a production task. In the case of the Mandarin speakers, subjects came from either mainland China or Taiwan, and neither this distinction nor any other dialectal information about the speakers was retained. Speakers were simply asked whether they were native speakers of Mandarin who learned Mandarin as their first language, and did not learn another language until later in life. All Mandarin speakers spoke English as well, though with varying degrees of proficiency. All English speakers were speakers of American English, from diverse locations throughout the US. 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.

2. Stimuli

a. Design and creation of stimuli. In order to create brief f0 samples from various locations in speakers' ranges (to be presented to listeners later), both groups of speakers were asked to perform two tasks. The first task was designed to estimate the speakers' individual f0 ranges, using a method common for clinical or experimental purposes (Reich et al., 1990; Zraick et al., 2000; Honorof & Whalen, 2005). Speakers were asked to produce rising or falling glissandos, described to them as rising or falling "pitch sweeps", using the vowel /ɑ/. First, rising sweeps through a speaker's range were elicited by asking speakers to start at a comfortable speaking pitch and to perform a rising sweep up to their highest achievable pitch, and they were aided by a rising sinusoidal tone presented over an earbud. In a second sweep speakers did the same, but instead started from a midpoint in their range and produced a falling sweep down to their lowest achievable pitch. Subjects were also instructed to speak rather than sing these sweeps through their ranges, and were permitted to practice this several times before being recorded. After the speaker and the experimenter determined that sufficient practice had taken place, the speaker was recorded performing the rising and falling sweep three times each. Recordings took place in a sound-attenuated booth using a Shure SM10A head-mounted microphone, digitized at 44.1 kHz by an XAudio A/D box with PCQuirer, and later converted to WAV files.

The recordings made of speakers in the glissando task were used to determine speakers' individual f0 ranges as follows. Inspection of f0 tracks using the autocorrelation method in Praat (Boersma & Weenink 2008) was used to locate the upper and lower limits of the speakers' physiological ranges. Of the repetitions of sweeps recorded for each speaker, the top of the range was defined as the highest f0 a speaker could sustain, usually including some f0s which would likely be considered to fall within the falsetto register; the floor of a speaker's range was defined as the lowest f0 reached that was analyzable by Praat's autocorrelation. From these values, the highest and lowest achievable f0s were retained for each speaker, and the range was calculated as the difference between these two points.

The second task was designed to elicit level f0 samples, from which brief, steady-state tokens from various locations in speakers' own ranges could be extracted. In order to collect such samples, speakers were asked to produce a series of spoken /ɑ/ vowels at various level pitches in their range, described as "level steps" lasting approximately 3 to 4 seconds each. They first performed steps beginning from a comfortable midpoint in their range, progressively producing level tones at higher pitches until they could no longer sustain the f0. They then performed the same task, but from their comfortable midpoint to progressively lower f0s, until they reached a point at which they were no longer able to sustain phonation. The experimenter guided speakers in this task during a practice session. Again, after the experimenter and speaker had determined that sufficient practice had taken place, the speaker was recorded producing each series of steps three times. Recordings took place in a sound-attenuated booth and were digitized as above.

The level /ɑ/ steps recorded from the speakers were used to create briefer, 500 ms /ɑ/ tokens that would be presented to listeners in the experiments below. The 500 ms tokens were portions extracted from the recordings of the actual steps, using the first 500 ms duration of the step most free of f0 or amplitude excursions or perturbations. A 50 ms linear amplitude ramp was applied to the beginning and end of each extracted token in order to create stimuli with auditorily less abrupt onsets and offsets. After this was completed for each of the steps recorded from each speaker, the highest and lowest token recorded was selected, and the seven tokens most evenly spaced between those two f0s were selected. This resulted in 180 tokens (9 tokens ×

20 speakers) with f0s spaced as equally apart from each other as was possible given the range of steps elicited in the task.

The f0s for these tokens, and their locations in each speaker's range are shown graphically in Fig. 1. 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. This may be due to the task, as some researchers have argued that the step task systematically provides a more modest estimation of a speaker's maximum f0 (Reich et al., 1990; Baken and Orlikoff, 2000), although this is not an entirely consistent finding (Zraick et al., 2000). Unlike Honoroff and Whalen (2005), we did not first establish each speaker's range, and prompt the speakers to produce pitches spanning that range. Nonetheless, for each speaker, a substantial portion of their physiological range (and a much wider range than normal speaking f0) was represented by the nine roughly equally-spaced tokens that were to be presented to listeners, who would be asked to determine where in an individual speaker's range a token came from.

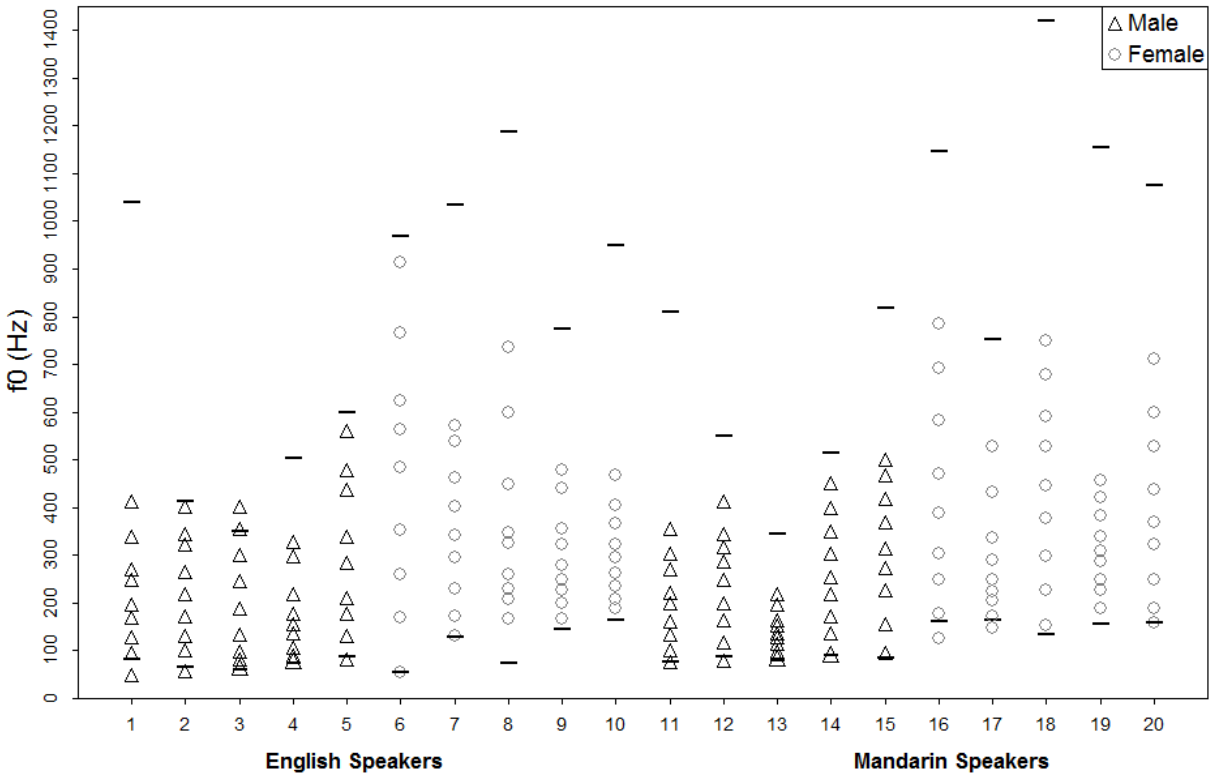

FIG. 1. Range of tokens recorded from English (1-10) and Mandarin (11-20) speakers in the step task, used as stimuli in the perception experiment. Male speakers are represented by dark triangles, female speakers by light circles.

Finally, in addition to the stimuli created from the twenty speakers' productions, a set of synthetic tone stimuli was also created, the purpose of which was to allow for the investigation of listeners' use of f0 itself to judge f0 "location" in a hypothetical voice when no other acoustic properties of a voice were available. Specifically, ten level sawtooth tones were created in Audacity, ranging, in 50 Hz intervals, from 50 Hz to 950 Hz. These tokens were also made to be 500 ms in duration, and were given linear amplitude ramps of 50 ms at onsets and offsets, as was done for the speech stimuli.

b.Acoustic properties of the stimuli. In order to determine the acoustic properties of the stimuli apart from f0, a number of measures were collected (for voice, not synthetic stimuli), for all 180 tokens. 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. 2009), which automatically collected several measures reflecting characteristics of voice. These included cepstral peak prominence (CPP, Hillenbrand et al., 1994) and the relative amplitudes of the first and second harmonic (H1*-H2*); both measures have been shown relevant to perceived breathiness in linguistic (Esposito, 2010) and non-linguistic (Klatt and Klatt, 1990; Hillenbrand et al., 1994; Hillenbrand and Houde, 1996) tasks. Also collected were measures of the amplitudes of H1 relative to that of the first and third formants (H1*-A1* and H1*-A3*, respectively). 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). Where a measure includes a harmonic amplitude, the values represent corrections made for formant frequency and estimated bandwidth (Hanson, 1997; Hanson and Chuang, 1999), as indicated by asterisks. The final measure of voice quality we included, the amplitude of H2 relative to that of the fourth harmonic, 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 a wide variety of measures of the glottal source spectrum. Seventy-eight such measures were made from seventy voice samples; principal components analysis of the 19 measures from the spectrum of the full audio signal indicated that four of them accounted for 76.6% of the variance in the measures. The fourth of these four factors was associated (only) with H2-H4 and accounted for 8.3% of the total variance in the measures. The first through third factors were associated with H1-H2, overall spectral slope, and high-frequency noise excitation, respectively. Thus H2-H4 represents some aspect of voice quality that is distinct from these other, more familiar measures. H2-H4 was also included in the set of measures applied to linguistic breathy versus modal voice qualities in Esposito (2010). However, this measure distinguished the two phonation types in only four of the ten language samples tested, and never uniquely, making it one of the least informative measures tested. Taken together, then, the results of these two studies suggested that H2-H4 captures some important aspect of individual voice quality that is possibly not exploited linguistically.

Subsequent exploratory work in our lab (Keating p.c., Garellek & Ward p.c.) compared the measures made by VoiceSauce across the voice samples produced by John Laver for the recordings accompanying Laver (1980). This comparison showed that the samples characterized as any variety of “falsetto” were distinguished primarily by their values on H2-H4, while no other voice quality was distinguished by this measure. Falsetto voice has a very high uncorrected H2-H4 and a very low (zero) corrected H2*-H4*. Furthermore, other recent work in our lab suggests that lexical tones (which differ primarily in f0) can be distinguished by this measure in Yi (Kuang, 2010), though not in Hmong (Esposito et al., 2009).

In sum, these results suggest that extreme values of H2-H4 indicate a voice quality with a high pitch, possibly characteristic of falsetto. This voice quality is statistically distinct from other voice qualities and is likely to be an important property of voices (Kreiman et al., 2007), it is not generally used for linguistic contrasts (Esposito, 2010), but it can distinguish tones (Kuang, 2010). Finally, it is important to note that, because the frequencies of the harmonics whose

amplitudes comprise this measure are twice and four times the fundamental frequency, H2-H4 is especially sensitive to the influence of the formant frequencies, even with low vowels such as the one used in our stimuli[1]. Therefore, corrections for the influences of the formants are especially important for this measure if f0 varies widely within or across speakers; for this reason all measures in our analysis are corrected, although because our stimuli were all produced with a low vowel, the other measures (e.g., H1-H2) would pose comparatively little concern. For all of the acoustic measures extracted by VoiceSauce, the mean value over the token was calculated and that value was used for analysis.

3. Listeners

20 native speakers of American English (mostly from California) and 21 native speakers of Mandarin (either Mainland or Taiwanese) participated as listeners in an f0-location rating task. None had participated as speakers in the production task described above. All listeners confirmed that they considered either English or Mandarin to be their native language. However, most being university students in the US, Mandarin speakers were bilingual in Mandarin and English (with varying levels of proficiency in English). All participants confirmed that they were given no previous diagnosis of a communication disorder and, to the best of their knowledge, had normal hearing.

4. Procedure

Listeners were presented with the steady-state /ɑ/ tokens taken from nine points at different location in speakers' f0 ranges. The stimuli were presented in two blocks, one containing English voices and one Mandarin voices, and ordering of the blocks was counterbalanced. 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. Tokens within each language block were randomized for each participant. Stimuli were presented to listeners at a comfortable listening volume (held constant across listeners) over Sony MDR V500 close, dynamic headphones 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 separately from voices), listeners were told that they would hear computer-generated tones at different pitches, and that for each tone they should rate how high or low it sounded compared to a human voice. More specifically, listeners were told to think of how high or low, in their experience, people's voices are, and to rate where the tone fell in that range. Responses were collected by way of 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 to slide an icon

[1] To see this, suppose that the first two formant frequencies for /ɑ/ are at 800 and 1200 Hz. Then, when f0 approaches 200 Hz, H4 will be boosted by F1. When f0 is around 300 Hz, H4 will be boosted by F2; H2 may be boosted somewhat by F1 but not as much. Thus in this range of f0, uncorrected values of H2-H4 are likely to be zero or negative regardless of the source spectrum. Conversely, when f0 is around 400 Hz, H2 will be boosted by F1, and when F0 is around 600 Hz, H2 will be boosted by F2; H4 may be somewhat boosted by F3. Thus in this range of f0, uncorrected values of H2-H4 are likely to be positive regardless of the source spectrum.

along a horizontal continuum. This bar was actually a scale from 0 to 100 that provided numerical output from the button's location on the scale for analysis. However, a continuous bar rather than this numerical scale appeared to participants, and they were not asked to think in terms of a numerical scale at any point in the study. Rather, participants were told that this continuum represented the speaker's pitch range, and, for each token, to slide the icon to the position in that range that it came from. The left edge of the continuum represented the very lowest the speaker could produce a pitch, and the right edge represented the very highest pitch the speaker could produce, and these ends of the continuum were labeled accordingly 'lowest' and 'highest'. Listeners were asked to listen to the token as needed, and then make a decision as to the location of the token in the speaker's range (rather than to listen to the token, move the button in the continuum, listen again, adjust the position of the button, and so on). Instructions were given to speakers in their own native language by a native speaking English or Mandarin experimental assistant. Participants were given a practice trial with three non-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 range, and (b) the f0 of the token (in Hz). Both of these correlations are shown in Fig. 2. In order to fit a logarithmic function to the group responses as a whole, tokens falling below a speaker's floor (as determined from the sweep task) were dropped from the correlation. This amounted to thirteen dropped tokens, all the lowest in a speakers' range.

Considering first the correlation between judgments of f0 location and the location of the token in the speakers' own ranges, we find the same relationship reported by Honorof and Whalen (2005); the best fit line indicates that the f0 location in range of the tokens accounts for 62% of the variance in the averaged listener ratings of the tokens (R^2 = .622). Unlike Honorof and Whalen, however, we found the correlation between f0 location and listeners' judgments of f0 location to be somewhat stronger when the sexes are considered individually (R^2 = .7318 for male voices, R^2 = .826 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, 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 more like 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.

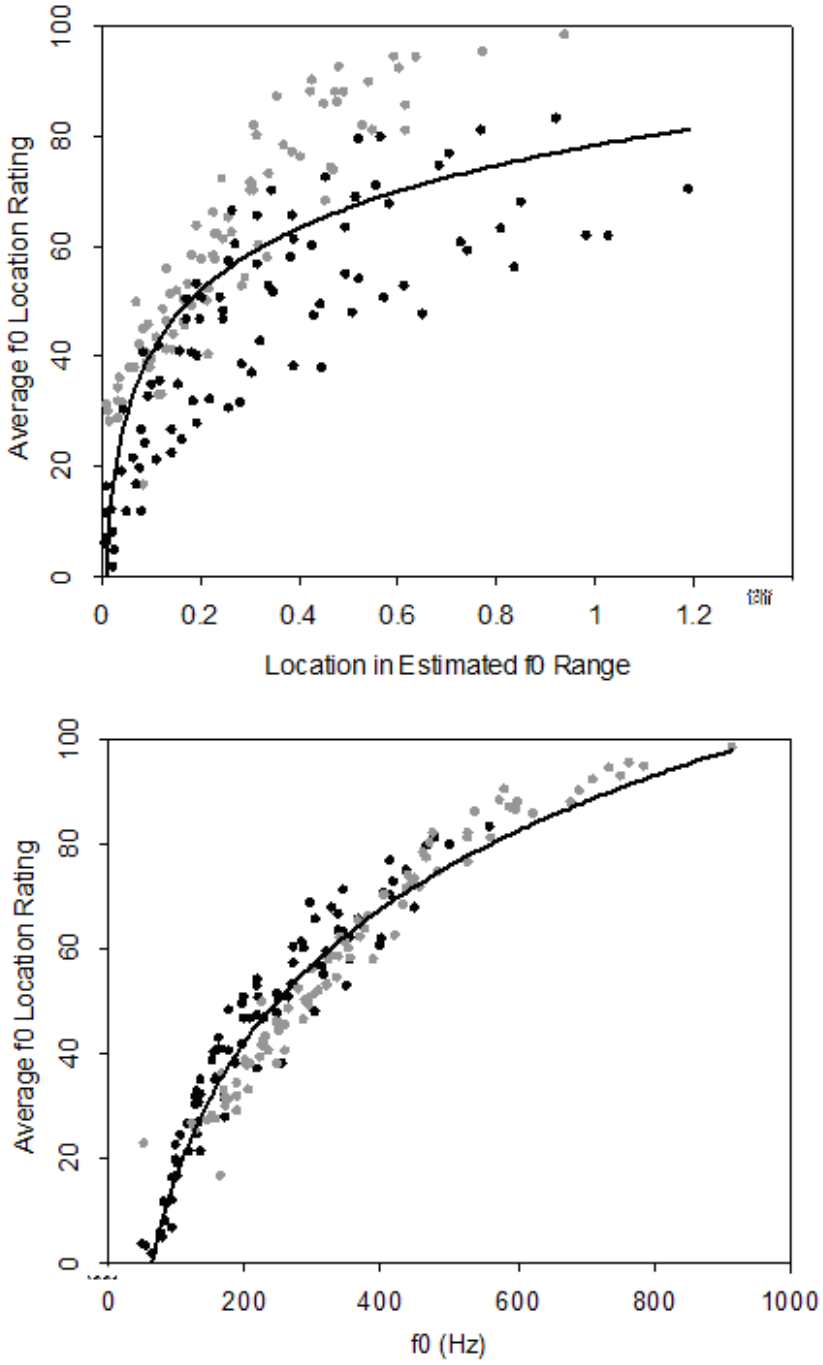

FIG. 2. Scatter plots showing averaged (pooled across listeners) f0 location ratings as a function of the f0 location of the tokens in the speakers' individual ranges (Top) and as a function of the absolute f0 of the token (Bottom). In both plots, tokens taken from the voice of male speakers are shaded in black, those from female speakers in gray; lines are fit to the group as a whole, however, both sexes combined.

The pattern is better understood when listeners' average judgments are plotted against absolute f0 rather than f0 location. As can be seen in the second correlation in Fig. 2, listeners' judgments of f0 location show a much stronger relationship with f0, the f0 of the tokens accounting for 93% of the variance in the averaged listener ratings of the tokens (R^2 = .933). The relationship was very similar when the sexes were considered separately (R^2=.942 for male voices, R^2= .927 for female voices). This would seem to suggest that the correlation between listeners' ratings of f0 location and actual f0 location was largely due to the necessary correlation between f0 location and f0, and that f0 was the better predictor. This can also be seen when comparing the sexes; a male token at a given f0 was rated higher by listeners than a female token at the same f0. This is what would be expected if listeners had the knowledge that, in fact, 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 location Z in range of a female should have a higher f0 than a token at the same location in a male voice; if listeners are making their decisions based primarily on f0, we expect female tokens. As just shown, that was in fact the pattern.

2. Modeling Listeners' Judgments of f0 Location

The correlations suggest that, at least when listeners are considered as a group, relatively little variance is left unaccounted for by f0. Nonetheless, it may be that factors such as voice quality contribute significantly to this remaining variance. Possibly such factors may be most useful when the f0 is within its most ambiguous region, or is atypical of the sex of the speaker. To determine which of many possible acoustic properties of the stimuli could have served to influence listeners' responses in addition to f0 in the f0 location rating task, those responses were modeled using linear regression, including both fixed and random effects. In particular, we modeled the outcome "rating" (which, as described earlier, was a value from 0 to 100), using speaker and item as random effects. The following fixed-effects parameters were included in the model: (a) the language of the listener (English or Mandarin); (b) the sex of the speaker (male or female), which here is a placeholder for whatever information listeners could use to decide this; (c) the f0 of the token; (d) the mean frequency of each of the first three formants (F1, F2, F3); (e) measures of voice quality: CPP, H1*-A1*, H1*-A3*, H1*-H2*, H2*-H4*. Among these fixed-effect parameters, the five measures of voice quality were permitted to interact (separately) with f0, listener language, and speaker sex; f0 was also permitted to interact with listener language and speaker sex, and listener language and speaker sex interacted in the model as well. More complex interactions were not included in the model.

To determine which of these fixed-effects parameters contributed to the most successful model of the outcome variable, we used a backwards process of elimination of parameters, comparing a full model that contained all fixed effects terms described above, with a series of sub-models that lacked one or more of the parameters. A log-likelihood ratio test (Baayen, 2008) was used to determine that the fit of a sub-model to the data was not equal or greater to the fit of the full-model; where the fit of a sub-model was, that simpler model was regarded as superior.

Using this method to remove non-significantly contributing parameters resulted in a model containing the following fixed-effects terms: f0; speaker sex; CPP; the frequencies of each of the first three formants; H1-A3; listener language; the interactions of f0 with speaker sex, CPP, F2, F3, and with H1-A3; the interactions of speaker sex with CPP, F2, F3, H1-A3, and with listener language.

Although all of these parameters contributed to the success of the model in a statistically significant way, it is highly unlikely that each contributed *equally* to the model. A process of model comparison, analogous to the one used to determine the best-fitting model, can also be used to compare the relative contribution of each of the variables within that model. The importance of a parameter to model fit is then indicated by the amount of improvement (again, measured by the increase in the log-likelihood estimate) the full model has over a model that lacks just that variable. To determine the relative weight of the fixed-effects parameters just listed, such a comparison was carried out so as to rank them according to their importance to the model, and is presented in Fig. 3. Because f0 had by far the greatest influence on the fit of the model, it cannot be included in the figure, else the scale would be made illegible. The remaining parameters are the sex of the speaker, CPP, the interaction of f0 and speaker sex, and F2. Because the individual contributions of each of the other factors in the model were comparably smaller and more incremental than f0, speaker sex (and their interaction), and H2*-H4*, we limit our discussion to effects pertaining to these four most influential parameters.

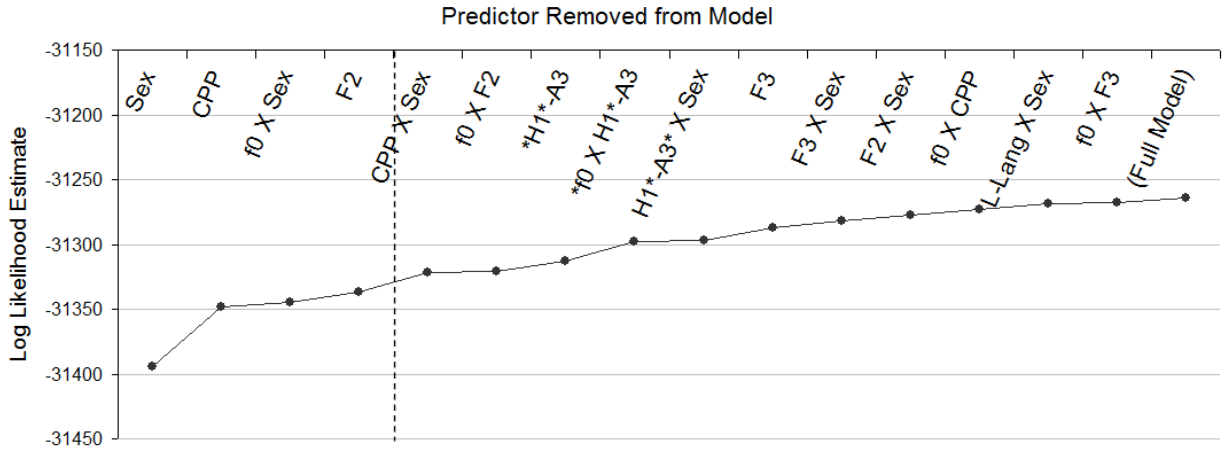

FIG. 3. Line chart showing the relative contribution of each parameter in the best-fitting model. The importance of a parameter is indicated by the difference in the log-likelihood estimate between a model lacking that parameter and the full model (model lacking no parameters). The largest difference was between a model lacking f0 (log-likelihood -33045, not shown), followed by a model lacking either the sex of the speaker (Sex), CPP, the interaction of f0 and speaker sex, or the frequency of the second formant.

The results of the mixed-effect linear model showed a significant main effect of f0 on listeners' ratings of f0 location of the tokens ($t = 5.63$, $p < .0001$). There was also a significant main effect for speaker sex ($t = 9.791$, $p < .0001$). Furthermore, the interaction of f0 and speaker sex was significant ($t = 12.91$, $p < .0001$). This interaction is shown in Fig. 4. In general, an f0 above approximately 250 Hz was associated with a higher f0 location rating when it came from the voice of a male speaker. The effect of F2 was also significant ($t = 2.67$, $p < .01$), and is also shown in Fig. 4 as a function of f0, although the interaction of the two was not one of the influential parameters in the model. In general, a token was judged as coming from a higher portion of a speakers' range (most apparent for higher f0s in the figure) if the F2 for the token was relatively low. Finally, although CPP was found to be an important factor in the model, the

main effect for this measure was not significant. This situation arises when a factor enters into significant interactions with very small effect size.

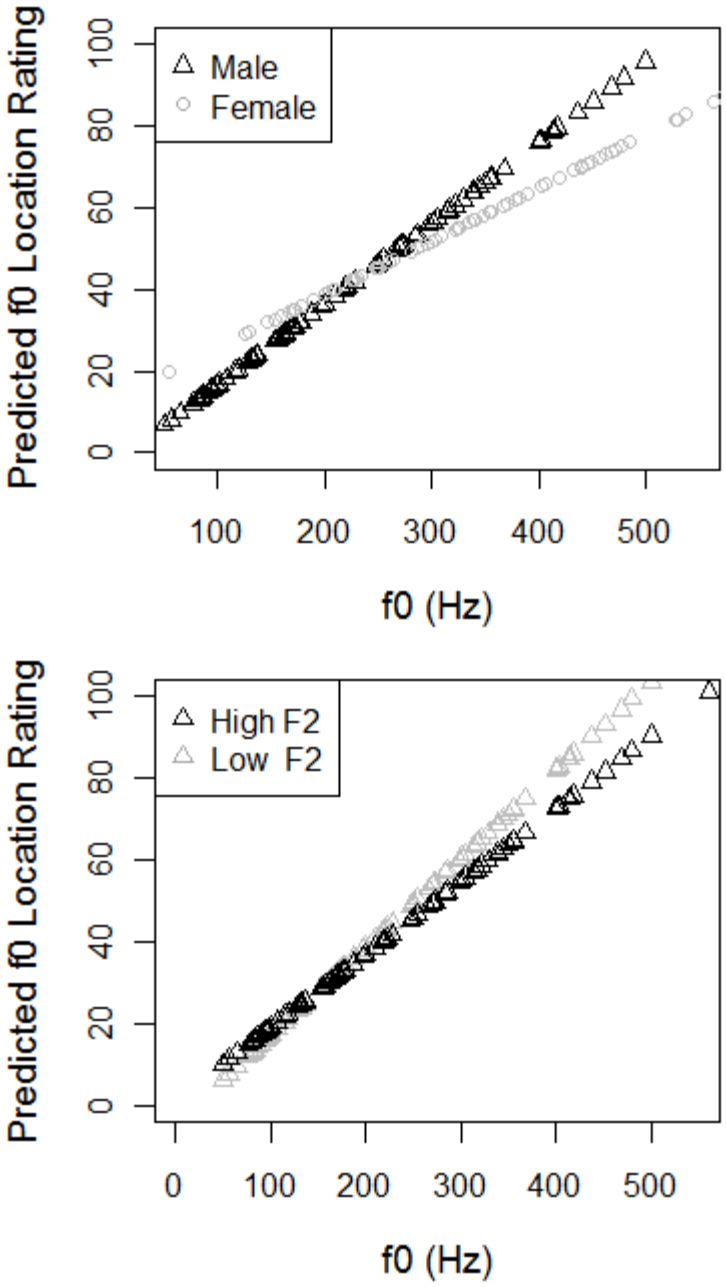

FIG. 4. (Top) Model predictions for f0 location ratings for male and female tokens, plotted as a function of f0; f0 values above the group mean (297 Hz) are associated with higher f0 location rating values when the voice is male rather than female. (Bottom) Model Predictions for f0 location ratings as a function of f0 at two different levels of the parameter “F2”. The high value represents F2 values one standard deviation above the group mean, the “Low” value one standard deviation below the group mean.

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 correlations for listeners' rating of the synthetic stimuli were quite similar (R^2 = .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 responses was carried out, as these synthetic tones possessed none of the other properties of voices of interest.

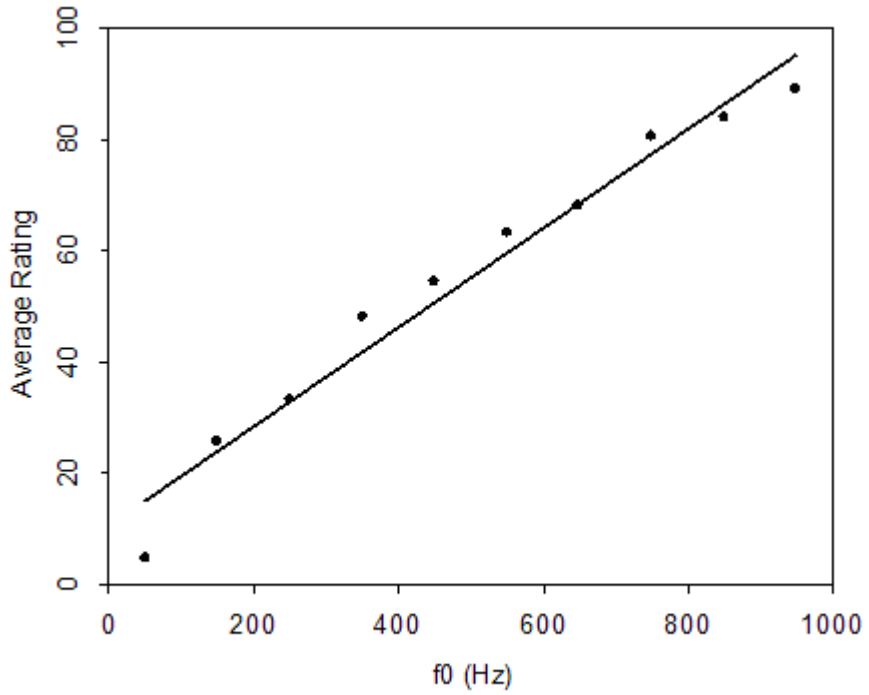

FIG. 5. Scatter plot showing averaged (pooled over listeners) f0 "location" ratings as a function of the f0 of each of the nine synthetic tone stimuli.

C. Discussion

In their correlational study, Honorof and Whalen (2005) found that listeners' judgments regarding the location of an f0 in the f0 range of a particular unfamiliar voice were correlated with the actual location of that f0 in that speaker's specific range. The question we sought to answer in Experiment 1 was what information in the signal listeners relied upon to make those judgments. One hypothesis of particular interest to us was proposed by Honorof and Whalen, and regarded listeners' possible use of cues to voice quality. Since various measures of voice quality are known to correlate with a speaker's f0, it is possible that it was also informative about a speaker's overall range.

The results of our model, however, which included a number of relevant measures, did not provide evidence for a major contribution by voice quality to listeners' judgments of f0 location. This is not to say that no role at all was played by voice quality, as CPP and H1-A3

both contributed to model fit in a statistically significant way. However, their contributions, indeed the contributions of all factors except f0 and the sex of speaker, appeared rather marginal—more of a 'fine-tuning' of the model's fit to the data compared with the two dominant factors. The model's results in this respect are not surprising given the correlations we found; the f0 of the tokens accounted for 93% of the variance in listeners' ratings of the stimuli, leaving very little for other measures, such as those of voice quality, and even speaker sex, to predict.

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, she does not really place it in the range of that speaker; rather, it is an idealized speaker, presumably one constructed from experience with many speakers, that is being judged. That listeners have quite acute expectations about how certain f0s relate to speakers' ranges was also shown very clearly by listeners' ratings of the non-voice tone stimuli. Although tones were presented in random order, listeners ranked the stimuli in a very linear manner, and their behavior in this task suggests strongly that listeners treat f0s systematically when f0 is the only human-like signal-intrinsic cue available. Our conclusion is that listeners' ratings of f0 location in our study, and also in Honorof and Whalen's (2005) study, had very little to do with the individual ranges of the speakers presented to them.

If listeners so clearly rely upon experience-based knowledge of generalized speakers' f0 ranges, the next question to ask is whether there are separate expectations for male voices and female voices, which would be likely since male and female speakers have characteristically different (but partially overlapping) ranges, at least for speaking f0s. Our results would seem to indicate that they do; the second most important factor in the model was the sex of the speaker, and its interaction with f0 was also a heavily weighted (relative to lesser factors) and highly significant factor. The predictions of the model suggested that listeners treated certain f0s (those 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 by a female. Again, this was also evident in the correlation of responses with f0 of the stimuli in Fig. 2.

This result also suggests that listeners were in fact able to make systematic decisions about the sex of speakers first, although we do not yet know how they might have accomplished this given the stimuli presented to them. That is, 'speaker sex' was included in the model, but speakers were not told which voices were male and which were female, making the factor a sort of cover term for some aspect(s) of the signal. One of the questions that needs to be answered, then, is what information listeners could have used to determine the sex of the speakers. As discussed in Section II, there are various contributers to the perception of speaker sex, including f0 and vocal tract resonances, and potentially voice quality. However, we have not established what aspects of the signal could be relevant to listeners' decisions as to the sex of the speakers used in Experiment 1, nor do we know to what extent it led them to the correct answers. It is particularly unclear how well these particular listeners actually judged speaker sex for the stimuli we presented to them in Experiment 1, because, as Honorof and Whalen (2005) note, unlike stimuli used in most studies of sex/gender perception, our stimuli were very brief and came from multiple speakers. Perhaps even more importantly, the tokens we presented to listeners came from a wide range of f0s. In Experiment 2 we investigated both listeners' accuracy in identifying the sex of the speakers from the stimuli used in Experiment 1, and, by way of modeling, we also explored the acoustic cues that serve as predictors of their judgments of a speaker's sex for this sort of stimuli.

III. EXPERIMENT 2

A. Methods

1. Stimuli

The stimuli were the same (voice only, not synthetic) stimuli presented to listeners in Experiment 1.

2. Listeners

23 native speakers of American English (mostly from California) and 23 speakers of Mandarin (either mainland or Taiwanese) participated in a sex identification experiment. None had participated in Experiment 1. All listeners confirmed that they considered English or Mandarin to be their first language, although most being university students, Mandarin speakers were bilingual (at likely a wide range of proficiency) in Mandarin and English. All participants stated that they had normal hearing.

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 position on along a continuum. For each voice token presented, listeners were to select one of two buttons, one labeled 'Male' and one female '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 in Mandarin and English, this simple task was effectively explained in English. 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), and participants were also 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, we first calculated the average probability of a correct response over each speaker's range of tokens. Collapsed over all speaker and listener groups, accuracy was on average 77.7% (SD = 28). We also submitted these f0 range-pooled scores to a three-way analysis of variance (ANOVA), each with two levels: the between-subjects factor listener language (English, Mandarin), and the two within-subjects factors speaker language (English, Mandarin) and speaker sex (male, female). The results of the ANOVA showed a number of significant effects and interactions. A significant main effect was found for speaker language [$F(1,44) = 50.15$, $p < .0001$]; 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 [$F(1,44) = 14.48$, $p < .001$]; 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 [$F(1,44) = 27.01$, $p < .0001$] and with speaker sex [$F(1,44) = 62.77$, $p < .001$]. In particular, both English and Mandarin listeners identified the sex of voices best when those

voices were English. A Bonferroni pairwise comparison (α = .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%, 13.4). Second, although there were no pronounced differences between the identification of English male (80.8%, SD = 12.2) and English female voices (79.1%, SD = 10.7), there was a significant difference between Mandarin male (66%, SD = 9.7), and Mandarin female voices (84.8%, 8.6).

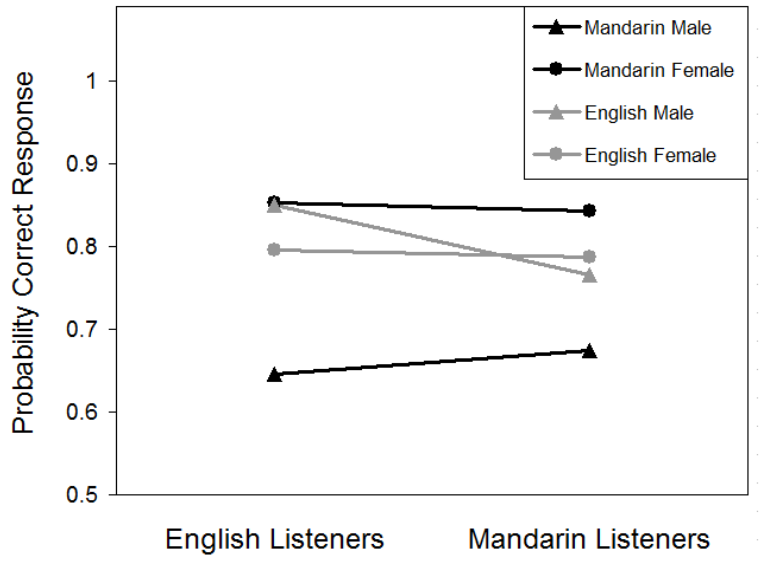

FIG. 6. Interaction plot showing the probability of accurate identification of speaker sex for the four speaker groups (pooled over the nine tokens of each speaker's range) by the two listener groups.

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 [$F(1,44) = 6.56$, $p < .05$]. 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). Bonferroni pairwise comparisons limited to the four speaker groups across the two listener groups, however, found the only significant difference to be for male English voices. 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).

2. Correlations with f0

Whereas the ANOVA allowed for a picture of how listeners responded to the speakers' voices overall throughout their range of f0s, previous research would lead us to suspect that accuracy would differ throughout a range of f0s, likely being lower when a speaker was outside of the range typical of the speaking f0s for 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, there is a close relationship between f0 and accuracy for all four groups. In general, female voices, especially Mandarin female voices, were by and large most accurately identified as female at f0 values above 200Hz, accounting for considerable portions of the variance in listeners' accuracy (R^2 = .38 for English female voices; R^2 = .43 for Mandarin female voices). Conversely, male voices were most accurately identified by listeners when f0 was below 200Hz, accounting for a much larger portion of the variance in accuracy for these groups (R^2 = .84 for English Males; R^2 = .73 for Mandarin Males).

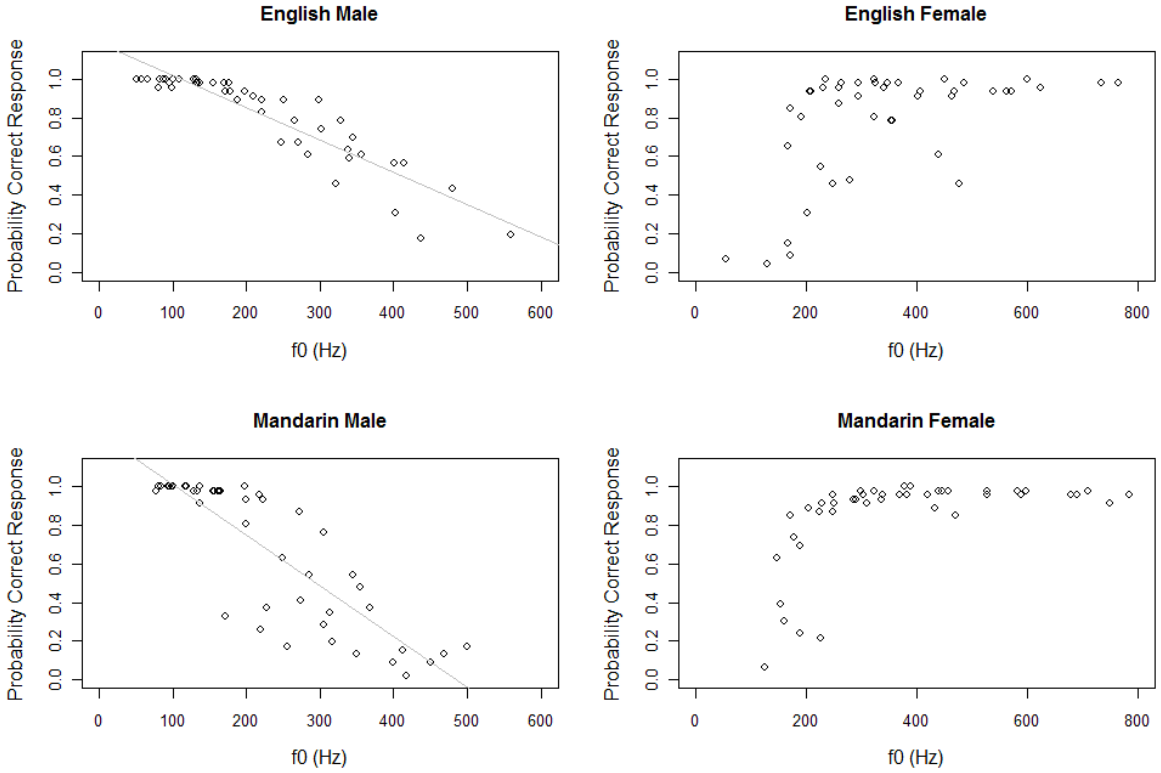

FIG. 7. Probability of correct identification of speaker sex as a function of f0 for the four speaker groups: English males (R^2=.84), Mandarin males (R^2=.73), English Females (R^2=.38), and Mandarin females (R^2=.43).

3. Modeling Listeners' Judgments of Speaker Sex

As was done to determine how listeners responded to the stimuli in Experiment 1, where the task was to judge location of a particular f0 within that speaker's range, we built a model of listeners' responses in the sex identification task to determine which acoustic factors served as predictors to those responses. In particular, we modeled the outcome "probability of male response" using logistic regression, again including both fixed and random effects. Random effects were, as in Experiment 1, speaker and item; fixed effects in the model were similar to those used in Experiment 1, and included the following: (a) the language of the listener (English or Mandarin); (b) actual f0; (c) resonances: the mean frequency of each of the first three formants; (d) measures of voice quality: CPP, H1*-A1*, H1*-A3*, H1*-H2*, H2*-H4*. Among the fixed-effect

parameters, the five measures of voice quality were permitted to interact (separately) with f0 and listener language, and listener language and f0 also interacted in the model.

To determine which of the fixed-effects parameters were actually relevant to the success of the model's fit to the data, we again used a backwards process of elimination of parameters, comparing a full model that contained all fixed effects terms, with a series of sub-models that lacked one or more of the terms. A log-likelihood ratio test was used to determine that the fit of a sub-model was not equal or greater to the fit of the full model, and where it was, the simpler model was selected as the best model. Using this method, the most successful model contained the following fixed-effects terms: 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.

A process of model comparison was again used to explore the relative contribution of each of the variables within the chosen model, the results of which are illustrated in Fig. 8. By far, f0 had the greatest influence on model fit, followed by F2. The next most important factors were H2*-H4*, 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 more incremental; we again limit our discussion to these five most important factors.

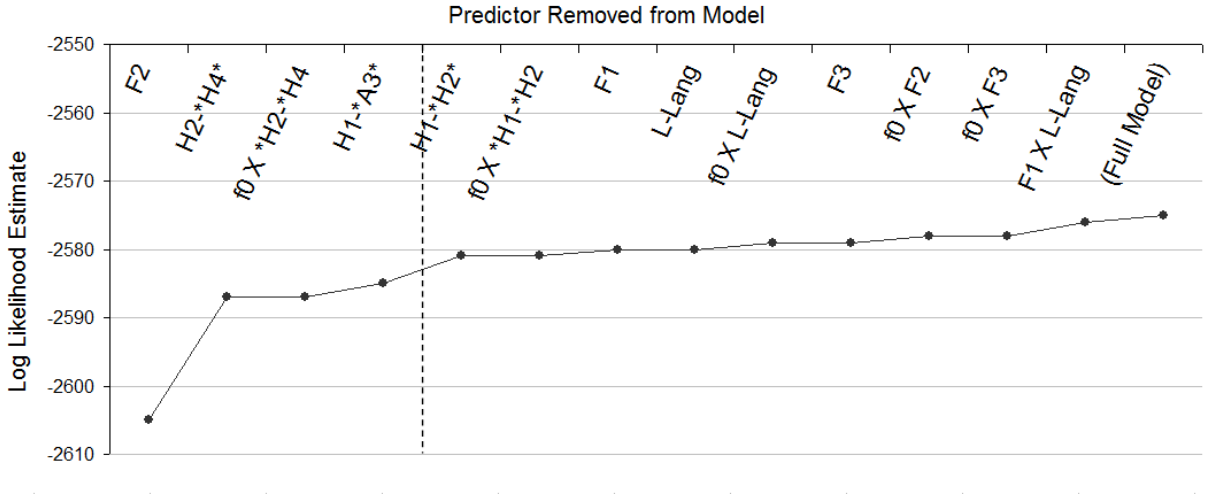

FIG. 8. Line chart showing the relative contribution of each parameter in the best-fitting model of 'male' responses in the sex identification task. The importance of a parameter is indicated by the difference in the log-likelihood estimate between a model lacking that parameter and the full model (model lacking no parameters). The largest difference was between a model lacking f0 (log-likelihood -3064, not shown), followed by a model lacking either F2, H2-H4, the interaction of f0 and H2-H4, and one lacking H1-A3.

The results of the model show a significant main effect of f0 on the probability of 'male' responses ($Pr(>|z|) < .0001$), such that higher f0s were strongly associated with a lower probability of male responses. Although F2 and H2*-H4* were both highly ranked parameters in the model, neither showed significant main effects (both ($Pr(>|z|) < .2$)). However, the next most important factor to the model involved H2*-H4* in interaction with f0, and this was significant ($Pr(>|z|) < .001$). The effect was such 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 by listeners. F2 also entered into an interaction with f0, and while this was much less important to

the model, it presumably accounted for F2's ranking in the model. The fifth and final parameter we consider here was H1*-A3*; it showed a significant main effect for H1*-A3*. When other covariates in the model are held at their mean values, lower values of H1*-A3* were associated with a higher probability of 'male' responses. These effects are plotted as a function of f0 in Fig. 9.

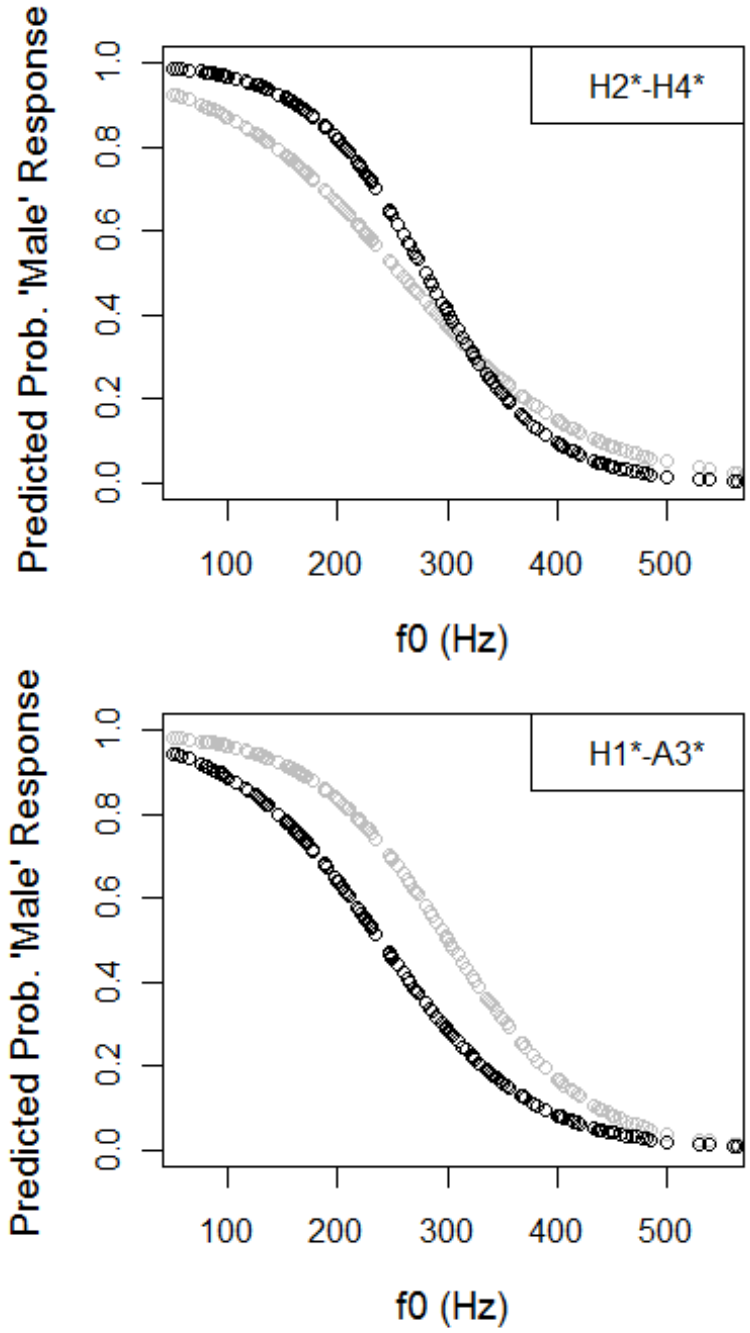

FIG. 9. Model predictions for probability of 'male' responses as a function of f0 for two levels of H2-H4 (top) and H1-A3 (bottom). The 'low' value for each parameter is one standard deviation below the group mean (gray); the 'high' one standard deviation above the group mean (black).

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 make obvious predictions about what characterizes a 'difficult' voice. We focus our discussion here on the least-well identified group of speakers: Mandarin males.

The first and most obvious explanation for why 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. This turns out to be consistent with the properties of these stimuli. Fig. 10 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, the Mandarin male speakers that were hardest for listeners to accurately identify as males 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 general theme of the both experiments presented above, which is listener attentiveness to primarily f0.

Clearly the relationship between f0 and listeners' judgments of the Mandarin male voices was not perfect, however; where it was not, we might assume that our other predictors can account for the gap. We therefore explored the Mandarin male stimuli further. We now wished to examine what acoustic properties other than f0 distinguished the most difficult Mandarin male from easier Mandarin male tokens. To do this, we calculated the mean values for a number of acoustic parameters (ones which factored into our models) between 150Hz and 350Hz for the Mandarin male stimuli. We identified which were 'easy' and which were 'difficult' tokens in this range by their position relative to the regression plotted for accuracy on Mandarin male tokens (plotted in Fig. 7): those above the regression line were relatively easier and those below were difficult. 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. Instead, we expect the two groups of tokens to be better distinguished by other factors in our model, in the general direction suggested by the model. Fig. 11 shows the measures that distinguish these two groups of tokens, and shows that the predictions are in several cases born out. Each of the measures shown in Fig. 11 suggests the tokens below the regression line had acoustic properties that would have led to their higher likelihood of being (wrongly) identified as female: F2 and F3, for example, were both higher. Similarly, H1*-A3* is noticeable higher for the Mandarin male tokens that were most difficult for listeners. 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.

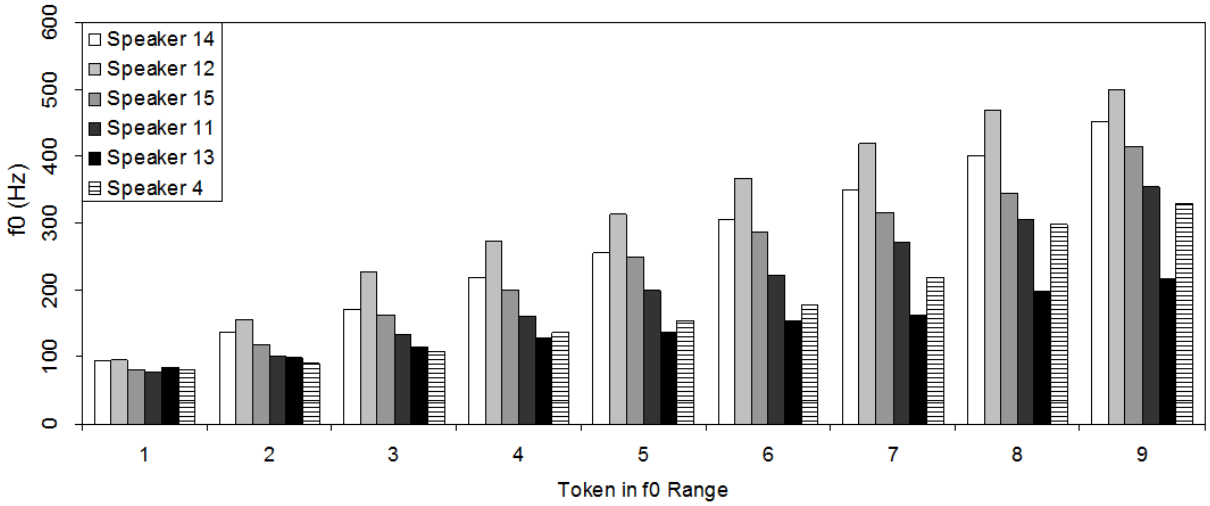

FIG. 10. F0s for each of the tokens from the five Mandarin male speakers, ordered from the least-well identified speaker (Speaker 14) to the best identified speaker (Speaker 13). Also shown is the best-identified English male speaker (Speaker 4). F0 is a good predictor of accurate gender identification for most tokens for most Mandarin male speakers.

C. Discussion

The purpose of Experiment 2 was to determine two aspects regarding the stimuli used in Experiment 1: how well listeners could have identified the sex of the speakers, and on what aspects of the signal those decisions were based. In terms of their accuracy, averages for all speaker groups were considerably above chance, although not the near-ceiling performance that many studies have reported. However, as noted earlier, the stimuli in used in our experiments were very brief, came from multiple speakers, and were produced over a wide range of f0s. One consistent finding in previous studies is that male voices are most difficult to identify when f0 is high for the average male speaking f0, and female voices are most difficult when the f0 is low for the average female speaking f0. Indeed, f0 has been regarded as the single most important factor in the perception of speaker sex, with resonance information about the size of the speaker's vocal tract being a close second. The accuracy results above are consistent with this in terms of f0. The effect of this tendency on accuracy was particularly evident for the English male voices, which showed a quite linearly-declining level of accuracy as f0 increased. In the case of female voices there was a less linear relationship between f0 and accuracy. If we assume an ambiguous region around approximately 200 Hz, this difference is expected given the range of the stimuli used in the experiment, however, as fewer female tokens would fall into the range of sex-ambiguous f0. While the male stimuli were more or less equally distributed between f0s below (43 tokens) and above (47 tokens) 200 Hz, the female voices used in Experiment 2 fell primarily above 200 Hz (15 tokens below, 75 tokens above; see Fig. 1). Thus if listeners are judging the sex of speakers primarily on the basis of the f0s which are typical of the average speaking f0s for the two sexes, the basic patterns in Fig. 7 are expected by virtue of the stimuli presented. This may be consistent with results in Lee (2009), where average male f0s for the high-onset tones were relatively high (near the f0 values of the female low-onset tones), and tones produced by males were not as accurately identified

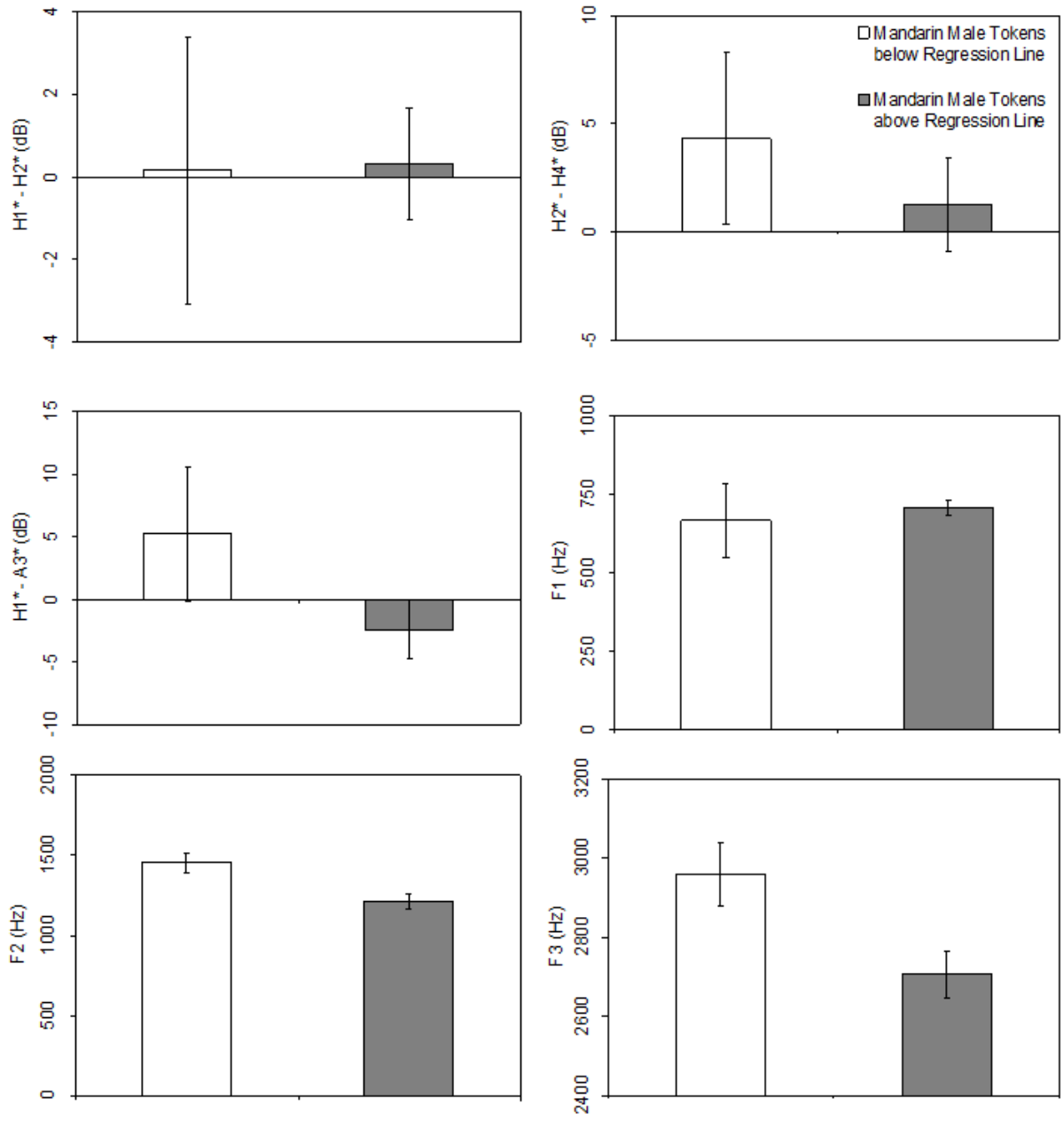

FIG. 11. Average values on six measures (H1*-H2*, H2*-H4*, H1*-A3, F1, F2, and F3) for Mandarin male tokens that fell either above or below the regression line for listeners' accuracy in sex identification. The tokens that fell below the regression line were more likely mis-identified by listeners than tokens above the regression, although all came from the same sub-range of f0s.

Indeed the results of our model, which was more acutely aimed at determining *how* listeners made their decisions, also confirmed the importance of f0. Other things being equal, an f0 below approximately 200 Hz was associated with a relatively high probability of a 'male' response by listeners, while that probability decreased as f0 increased. The second most important factor in listeners' judgments of sex was the frequency of F2, although we did not explore its role further because a main effect was not significant, and its interaction with other

factors in the models did not hold a primary position in the model. We note, however, that closer inspection of a subset of Mandarin male tokens indicated that one of the characteristics of mis-identified voices was a high F2.

Finally, there was an effect for the voice quality measures included in the model. Whereas previous production studies have found significant differences between males and females associated with voice quality (e.g., Klatt and Klatt, 1990; Hanson, 1997, Hanson and Chuang, 1999; Lee, 2009), these differences have been measures of breathiness (e.g., H1-H2, CPP), and spectral tilt (H1-A3), both generally showing higher values in female compared with male speakers. Our results above provide evidence that in fact H1*-A3* does reflect an aspect of the voice that listeners can use for the purpose of identifying an unfamiliar speaker's sex. As the production studies would predict, higher values for H1*-A3* are associated with female rather than male responses, and male speakers who do not fit this description seem to be less easily identified as male. Finally, higher values on measure H2*-H4*, not traditionally included in studies of voice quality, were associated with male responses, at least for a subset of f0s. In our model, H2*-H4* was the most important factor to the model after f0 and F2, and was highly significant, although it did not predict the properties of the subset of Mandarin male stimuli we explored. When we consider F2, H2*-H4* and H1*-A3*, however, it is important to keep in mind that their overall influence on the success of that model was quite modest compared to the most important factor, f0.

IV. GENERAL DISCUSSION

In this paper we sought to provide some evidence regarding how listeners in Honorof and Whalen's (2005) perception experiment managed to place an individual f0 within a speaker's range without any prior experience with that range, no syllable external information, and no dynamic syllable-internal f0 information on which to base a method of normalization. The apparent implication of their result was that listeners used other signal-intrinsic information to make decisions as to f0 location. One hypothesis we had a special interest in was that voice quality was one such source of information listeners could use for this purpose.

In addition, we also wished to clarify two different ways a method of normalization could utilize such cues. The first was in a direct way: listeners attend to acoustic parameter X and come to the table with experience-based knowledge that value Y for acoustic parameter X indicates location Z in a speaker's range. Indeed our interpretation of the results is that f0 was used in this way by listeners. By far the greatest predictor of listeners' judgments of f0 location was f0 itself. This is somewhat unsurprising for two reasons: first, f0 necessarily correlates highly with location in f0 range, so by judging a high f0 as relatively high, a listener is behaving reasonably. What this implies, however, is that listeners have expectations about f0s for *average speakers*. The clear interaction between f0 and speaker sex indicated that listeners in fact have separate expectations about f0 ranges for each of the sexes. This, too, is unsurprising; indeed, it would be surprising if listeners' previous experience with voices did *not* lead to some expectations about where a given f0 might fall within the range of an average speaker. We interpret the results for f0 to be indicative of those expectations.

Was voice quality also used in this direct sort of manner? In brief, no. We did not find that any measure of voice quality was a primary predictor of f0 location decisions. Note that this result is not necessarily surprising if voice quality and f0 are strongly correlated, as several such measures have been shown to be. As discussed at earlier, such a relationship renders voice

quality a redundant, and thus less informative, cue. Particularly since we have demonstrated that listeners have such salient expectations about speakers' likely ranges—and since f0 is a sufficiently salient auditory cue—it is reasonable to assume that listeners' use of other cues for the purpose of determining location in range will be relatively marginal.

Now let us consider the second, indirect method of using acoustic information to determine location in f0 range: listeners attend to acoustic parameter X, with the knowledge that value Y on acoustic parameter X informs them of the proper range to assume for the speaker. In this sense, signal-intrinsic information is used not to make a direct decision about the speakers' own individual range, but to assign the speaker to an individual or group range which is already known (based on experience and stored in memory).

This indirect use of acoustic information to judge f0 location was also evident; listeners' decisions about the location of an f0 in a speaker's range were partially dependent on the sex of the speaker, and decisions about the sex of the speaker were dependent on a number of acoustic parameters. Here f0 was shown again to be most relevant. Thus f0 was used both directly (listeners know what location in range a given f0 should belong to) and indirectly (by providing a basis for sex identification). Identifying which sort of role, if any, voice quality might have played was one of the primary goals of our study. Our conclusion is that its use to listeners in identifying location in f0 range is primarily in identifying a speaker's sex and, thus, is indirect. A number of studies cited earlier would have predicted it to be useful for this purpose, although we do not know of a study that has actually tested these cues, or their relative weights in a model. The measures we found to be most relevant were H1*-A3* and H2*-H4*. In this way we provide a specific statistical model that accords with Lee (2009) and Lee et al. (2010)'s hypothetical account of Honorof and Whalen (2005)'s findings.

That listeners associated higher values of H1*-A3* with female rather than male voices is one of the predictions previous studies would have made. H2*-H4*, however, has for the most part not played a role in the study of voice quality, and as a result it is unclear what property of the voice it reflects. As mentioned earlier, it may be characteristic of vocal production at high f0s, or possibly falsetto register. Although this requires further investigation, it appears that in our data, higher values were interpreted as more female-like, across the normal speaking f0 range for males and females. In contrast, at high f0s there is a trend for higher values to be interpreted as male. This could be because a male voice will have more "falsetto quality" at a high f0 than a female voice will, and listeners (with fairly vivid expectations about the f0 ranges of males and females) know how to interpret this. Although this scenario is highly speculative, our results suggest H2*-H4* plays a more prominent role in listeners' perception of speaker sex than any other, more common, measure of voice quality. Characterizing the aspect 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; and Jody Kreiman and Abeer Alwan for helpful discussions.

REFERENCES

Baayen, 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.

Boersma, P. and Weenink, D. (**2008**). Praat: doing phonetics by computer (Version 5.0.13).

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

Esposito, C., Ptacek, J., and Yang, S. (**2009**). "An acoustic and electroglottographic study of White Hmong phonation," J. Acoust. Soc. Am. **126,** 2223(A)

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

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

Henton, H., and Bladon, R. (**1985**). "Breathiness in normal female speech: Inefficiency versus desirability," Language and Communication **5**, 221-227.

Hillenbrand, J., Cleveland, R., and Erickson, R. (**1994**). "Acoustic Correlates of Breathy Vocal Quality," J. Speech Hear. Res. **37**, 769-778.

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

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

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

Klatt, D., and Klatt, L. (**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 Sidtis, D. (**2011**). *Voices and Listeners* (Wiley-Blackwell, Hoboken, NJ), Chap. 4, section 4.3.

Kuang, J. (**2010**). "An acoustic and electroglottographic study of phonation contrast in Yi," J. Acoust. Soc. Am. **127**, 2022(A).

Laver, J. (**1980**). *The phonetic description of voice quality* (Cambridge University Press, Cambridge). Audio cassette illustrations for Chapter 3.

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

Lee, C. (**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.

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

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

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

Reich, R., Frederickson, R., Mason, J., and Slauch, R. (**1990**). "Methodological variables affecting phonational frequency range in adults," J. Spinal Disord. **55**, 124-131.

Yen-Liang Shue (**2010**). *The Voice Source in Speech Production: Data, Analysis and Models*, unpublished doctoral dissertation, UCLA, Chap. 4, section 4.4.

Shue, Y., Keating, P., and Vicenik, C. (**2009**). "VOICESAUCE: A program for voice analysis," J. Acoust. Soc. Am. **126**, 2221(A).

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

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