Perception of pitch location within a speaker's F0 range[a] 1385

Douglas N. Honorof[b] and D. H. Whalen
Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06511

(Received 24 March 2004; revision received 2 November 2004; accepted 5 November 2004)

Fundamental frequency (F0) is used for many purposes in speech, but its linguistic significance is based on its relation to the speaker's range, not its absolute value. While it may be that listeners can gauge a specific pitch relative to a speaker's range by recognizing it from experience, whether they can do the same for an unfamiliar voice is an open question. The present experiment explored that question. Twenty native speakers of English (10 male, 10 female) produced the vowel /ɑ/ with a spoken (not sung) voice quality at varying pitches within their own ranges. Listeners then judged, without familiarization or context, where each isolated F0 lay within each speaker's range. Correlations were high both for the entire range (0.721) and for the range minus the extremes (0.609). Correlations were somewhat higher when the F0s were related to the range of all the speakers, either separated by sex (0.830) or pooled (0.848), but several factors discussed here may help account for this pattern. Regardless, the present data provide strong support for the hypothesis that listeners are able to locate an F0 reliably within a range without external context or prior exposure to a speaker's voice. © *2005 Acoustical Society of America.* [DOI: 10.1121/1.1841751]

PACS numbers: 43.71.An, 43.71.Bp, 43.71.Es [PFA]. Pages: 2193–2200

I. INTRODUCTION

Fundamental frequency (F0) carries information about many different linguistic and paralinguistic aspects of the speech signal, but it does so in a speaker-dependent way. A pitch-accent system like that of Japanese, for example, distinguishes words on the basis of which mora bears a shift in relative F0. However, F0 height is not specified by the language in general; rather, what counts as high or low varies by speaker (Leather, 1983; Moore and Jongman, 1997). Irrespective of which parameters require F0 manipulation in a given phonology, the most obvious source of difference in pitch is speaker sex. Peterson and Barney (1952) found a difference of approximately three-quarters of an octave between adult males and adult females in average F0s of citation-form monosyllables across vowels, and approximately a full octave's difference when comparing analogous F0 measures for adult males and children. It appears, then, that "high pitch" cannot be specified as a particular F0 value, because any value (if attainable) could be high for some speakers and low for others.

How, then, do listeners gauge where within the speaker's range a particular pitch lies? One possibility, of course, is that direct exposure to a voice over time allows a listener to learn the range of pitches that a speaker can and does produce. If so, the perception of linguistic pitch targets may become possible as a function of exposure to a speaker's voice. Indeed, there is evidence that familiarity with a voice is a factor in speech perception. For example, words spoken by familiar voices are easier to recognize (Palmeri *et al.*, 1993) and identify in noise (Nygaard and Pisoni, 1998). Such effects of familiarity in perception are not surprising given that linguistic and sociophonetic features are intertwined with speaker-indexical information in the speech signal (Ladefoged and Broadbent, 1957; Fellowes *et al.*, 1997; Remez *et al.*, 1997) and in word memory (Goldinger, 1996). It may be that listeners perform "speaker normalization" in order to strip away linguistically extraneous information from the linguistic message, but it is at least as likely, as recent findings suggest, that listeners use a normalization strategy in order to disentangle invariant phonological information and invariant speaker information from each other without necessarily forgetting the latter (see Sheffert and Fowler, 1995). There may remain much that we do not yet understand about how speaker normalization is accomplished by human listeners (e.g., Johnson and Mullennix, 1997), but we should be careful not to assume that perception always entails a computational process of normalization. Clearly, listeners successfully perceive the speech of unfamiliar voices—voices not stored in memory. In such cases, no exposure-based normalization strategy is available to the listener.

It is unclear whether previous exposure to a voice is required for accurate pitch perception. When a sentential context is given, listeners can estimate the location of a pitch within a speaker's range (Wong and Diehl, 2003). Other studies suggest success with contexts even shorter than a sentence. For example, in some languages, there are utterances as short as a single syllable that span a large region of the speaking range as in the case of the falling tone of standard Mandarin, in which case exposure to a single utterance of a single tone might, conceivably, provide enough information about a speaker's range to allow listeners to normalize F0. Indeed, a study by Moore and Jongman (1997) indicates that perception of a Mandarin tone having an F0 turning point or change in F0 does not depend on context, thus dynamic F0 is presumably calibrated syllable internally. If prior exposure to a short contour is required for the listener to succeed at normalizing pitch, logically, the listener should

[a]Portions of this work were presented at the 146th Meeting of the Acoustical Society of America, Austin, Texas, November 2003.
[b]Electronic mail: honorof@haskins.yale.edu

not be able to normalize pitch correctly if presented only with *steady* F0s. It remains to be seen whether listeners can correctly assess the location of F0 within the spoken pitch range (*tessiture*) of an unfamiliar individual on the basis of isolated *natural* utterances without first having been exposed to F0 contours within the speaker's range.

If perception of steady F0 does not require prior exposure to the speaker's range, we must determine how the listener is able to accomplish the task. One basis for pitch location within an F0 range might be voice quality. The literature has produced a number of differing claims regarding the nature of dependencies between voice quality (as reflected in glottal pulse characteristics) and F0. Common measures of glottal pulse characteristics include overall intensity, the open quotient of the voice source (that is, the relative time during which the glottis remains open during a pitch period), skewness of the glottal pulse, the strength of the relationship between the amplitudes of F0 and F0*2 (that is, H2–H1), etc. Swerts and Veldhuis (2001) review this literature and present the results of a highly constrained study of their own aimed at exploring the proposal that listeners "might be able to locate an utterance in a speaker's speech range on the basis of voice quality (298)." They conclude that there is some evidence that the perception of F0 depends on voice quality, but do not go so far as to explore directly the ramifications of such a finding for perception (see also Fant *et al.*, 1985; Klatt and Klatt, 1990; Gobl and Ní Chasaide, 1992). Di Paolo and Faber (1990) have reported that voice quality (breathy versus creaky phonation as indexed by F1–F0) can be used by listeners to distinguish tense from lax vowels. Taken as a whole, these studies make direct assessment of voice quality a possible means for the immediate normalization of an individual speaker's pitch without prior exposure to a broad range of productions by the voice of the individual. Therefore, we hypothesize that listeners should be able to use immediately available information about voice quality to locate where within a speaker's range the pitch of a vowel lies, even when presented with unfamiliar voices and virtually no context for their judgments.

If listeners are unable to perform this task, we will have to conclude that the perception of pitch does indeed require contextualization or familiarization with a particular speaker's pitch range. If, on the other hand, as predicted, listeners are able to successfully identify the relative location of isolated steady pitches without prior exposure to a speaker's voice (and thus to a speaker's tessiture), we will have evidence that acoustic parameters of some sort must co-vary with F0 so as to make an immediate identification of relative F0 possible.

II. METHODS

A. Materials

1. Speakers

Twenty native speakers of North American English (10 men, 10 women) participated in the production phase of the experiment. Ages ranged from 20 to 78. Each speaker passed a prescreening administered by questionnaire. The answers of those who participated in the production phase of the experiment indicated that they had neither training nor experience as a vocal soloist, that they were nonsmokers (for at least the preceding five years), that they were in good vocal health (including the absence of symptoms of congestion, coughing, post-nasal drip, or fatigue on the day of the test), that they were not prone to severe or chronic respiratory allergies, and that they had never received the diagnosis of a communication disorder, vocal-fold pathology, or neuromotor impairment. Furthermore, on the day of the test, all 20 speakers passed manual binaural audiometric screening via a Beltone 110 in a sound-attenuated chamber. The audiometric tests demonstrated hearing-threshold levels at audiometric (pulsed) frequencies between 0.125 and 1 kHz at 25 dB or lower—well below the average amplitude of conversational speech.

2. Stimulus generation

In order to determine which F0s we should attempt to elicit from a given speaker and in order to allow us, at a later date, to calculate where those F0s lay within speaker-specific voice ranges, we elicited spoken tessitures (vocal ranges) from each speaker. In a double-isolated sound booth, the speaker was instructed to speak, not sing (Titze, 1995), /ɑ/ glissandos as though they were saying the words at a conversational "loudness" to the experimenter, who was seated in the booth approximately one meter away from the speaker. The speaker spoke an /ɑ/ by beginning at a comfortable (self-determined) habitual pitch (see Fairbanks, 1960), then increasing fundamental frequency continuously until modal phonation could no longer be sustained. Multiple rising glissandos of this type were elicited from each speaker until the experimenter was satisfied that he had recorded one smoothly rising, nonsung glissando that included the extreme high end of the speaker's range. Next, a set of falling /ɑ/ glissandos was elicited in a similar manner beginning at a speaker-specific habitual pitch and falling until the speaker was no longer able to sustain phonation at the low end of the range. The number of attempts was typically three but varied from two to eight. During elicitation of spoken glissandos, the speaker was asked to begin again if he or she sensed, or the experimenter judged, that the voice had taken on a sung quality or if the audio signal was clipped or if a vowel token was produced noticeably more quietly than other tokens.

Speaker-specific ranges were calculated from these glissandos using autocorrelation in Praat v4.1 (range=45–900 Hz; 15 maximum candidates), in particular, to provide information about the high end of the tessiture. Autocorrelation was supplemented by interactive measurement of the reciprocal of the arithmetic mean of three successive pitch periods. The floor of the spoken tessiture was defined as the basal pitch, specifically, the frequency at which periodic F0 was no longer analyzable by autocorrelation or, in dubious cases, identified by direct measurement of three pitch periods. The ceiling of the spoken tessiture was defined as the frequency of the first falsetto break extracted from the rising glissando during which nonfalsetto voice quality was sustained at the highest frequency. In clinical practice, falsetto is often included in a normative voice range profile (VRP, also called the "$F_0 - SPL_{min}$ voice profile," "$F_0/_{SPL}$ profile,"

"voice field," "voice area," or "phonetogram") in an attempt to assess voice disorders (especially of singers), in which case the falsetto break is often seen as a small notch in intensity (<5 dB) on the loud side of the (normally) oval shape of the profile (Schutte and Seidner, 1983; Heylen *et al.*, 2002). Although falsetto may be used in speech, not just in singing, we excluded falsetto from our speaker-specific ranges in order to increase the difficulty of the ranking task for the listeners (see later). Because we were not interested in the margins of the F0 range, no effort was made to control for variations in extrema within speakers as a function of time of day (see Gelfer, 1989), beyond the collection of multiple ranges from each speaker.

Once speaker-specific ranges had been established, each speaker was asked to sustain a spoken /ɑ/ on at least eight distinct notes spaced roughly evenly within his or her nonfalsetto spoken F0 range. The instruction to speak rather than sing /ɑ/ was met with surprisingly little resistance, perhaps because nonsingers had been recruited by design, some of whom having expressed relief that they were not expected to sing. As was the case with the glissandos, occasional lapses into a sung voice quality also occurred in the production of steady tones, and sung tokens were replaced. In order to help these nonsingers space notes out within their own ranges, they were provided with sinusoidal audio stimuli. Specifically, they heard four 300 ms pure tones per octave (notes c, e, g, and a from 49 to 524 Hz) in which a linear ramp function was applied to the amplitude over the first 100 ms and over the last 100 ms. The experimenters selected at least eight such auditory stimuli within the speaker's tessiture and presented them over a harman/kardon multimedia speaker system with adjustable output levels. Output was always well above measured auditory threshold levels for the speaker. The talker was asked to respond to each stimulus by attempting to match the pitch of the tone while speaking /ɑ/ at a conversational "loudness." In order to help the speaker further, we used a Kay Visipitch II running in DOS to provide real-time visual feedback on success of pitch-matching. The participant saw a trace of his or her own pitch along with horizontal guide lines placed at 10 Hz above and 10 Hz below the target pitch. Speakers were asked to hold each tone for four to five seconds, then to inspire (to encourage resetting of the glottis) and to produce at least another two vowels, also separated by a breath. Vowels so produced were also digitized as 44.1 kHz, 16-bit NSP sound files using a near-measurement-quality microphone (Earthworks QTC1) routed to the Visipitch via a hardware high-pass filter (Focusrite ISA220, cutoff=30 Hz, 18 dB/octave rolloff) under speaker-independent input gain. We excluded tokens with noticeable F0 or intensity excursions, as seen in acoustic analysis displays. For each speaker, the eight tokens that were best spread out throughout the speaker's range were then selected. For each item selected, the first 500 ms containing no noticeable F0 or intensity excursions were extracted from the longer vowel. A linear amplitude ramp was applied to the first and last 100 ms of this extracted segment.

B. Listeners

Twelve native speakers of North American English (6 men, 6 women) participated in the listening task. Ages ranged from 20 to 58. Each listener was asked to respond to a prescreening questionnaire. Answers indicated that none had ever been diagnosed with a communication disorder. Furthermore, on the day of the test, all 12 listeners passed a manual binaural audiometric screening via a Beltone 110 in a sound-attenuated chamber. The audiometric tests demonstrated hearing threshold levels at audiometric (pulsed) frequencies between 0.125 and 1 kHz at 30 dB or lower, well below the lowest amplitude present in the stimulus materials.

C. Procedures

For each stimulus we had recorded, we calculated the achieved F0—*not* the target F0—as a percentage of the corresponding speaker-specific nonfalsetto tessiture using the formula in Eq. (1), where $F0_{ratio}$ is the percentage of the individual speaker's tessiture with which we aimed to correlate listener rankings, $F0_{ach}$ is the fundamental frequency location actually achieved by the speaker, T_{base} is the lowest sustainable F0 within the speaker's spoken tessiture, and T_{range} is the tessiture itself.

$$F0_{ratio}=\frac{F0_{ach}-T_{base}}{T_{range}\cdot 0.01}. \quad (1)$$

Tessitures were based on spoken glissandos, not steady F0s, but highest speakable steady F0s and the top end of the nonfalsetto spoken glissando ranges differed slightly, as reported elsewhere (Reich *et al.*, 1990, but see Zraick *et al.*, 2000). Therefore, for two female speakers, the highest F0 values lay above the speaker's tessiture, 3% and 14% above for one speaker, and 5% above for the other. Half of the lowest male tokens and half of the lowest female tokens, though they fell within the tessiture, were produced with highly periodic glottal fry. Two instances of the second lowest tones (one male, one female) were also produced with fry. Tessitures for each speaker appear coded by sex in Fig. 1.

Listeners were warned that they would be hearing the voices of different speakers using different parts of their speaking ranges, and asked to identify where within the speaker's range each token fell. The list of stimuli, consisting of one token of each of eight F0s collected from each speaker (8×20=160 stimuli), was randomized across speakers and tokens separately for each listener and presented diotically one at a time over headphones. Stimuli were presented at a comfortable output level set by the experimenter and kept constant across listeners. Responses were solicited via a custom MATLAB graphical user interface (GUI) on a flat-panel display. The GUI window accepted input via clickable buttons numbered sequentially and arranged vertically, with the word "high" at the top and "low" at the bottom. Listeners used these buttons to indicate the location of an /ɑ/ token within a speaker's pitch range on a nine-point scale numbered "1" ("just about as low as they could go") to "9" ("just about as high as they could go"). A nine-point scale was used so that the eight F0 levels would not be underrepresented. Listeners were instructed to choose "5" if they

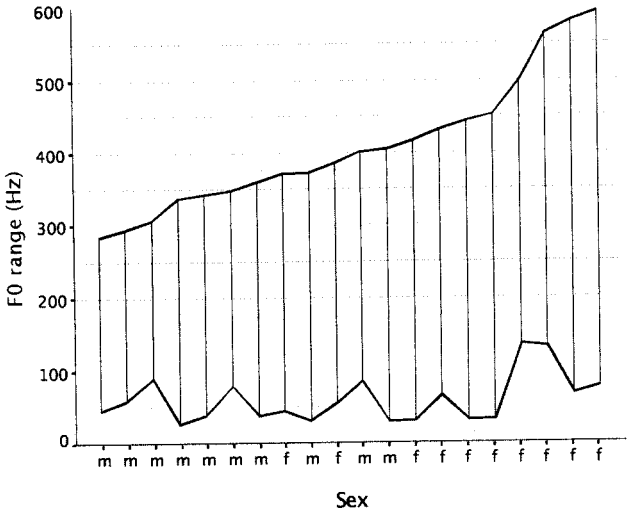

FIG. 1. Line chart of tessitures by speaker with high and low values connected. The chart is arranged from lowest falsetto break (left) to highest (right). Male falsetto breaks cluster toward the low end and gradually rise toward the female falsetto breaks with little overlap. Sex-based differences in the location of the low end of the tessiture are less striking. There, ranges bottom out in the high 20s and low 30s for the lowest voices in both sexes, though the 3 highest points are female (594, 563, 581, and 563 Hz) and are measured at over 150 Hz higher than the highest male low end.

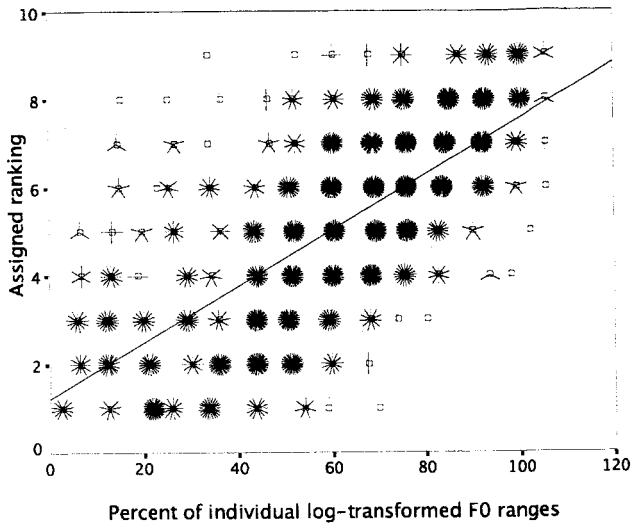

FIG. 2. Sunflower scatterplot with regression line between rankings assigned by 12 listeners and log base 10 transformed locations of eight spoken F0s per speaker (20 speakers). Locations are reported as a percentage of speaker-specific spoken F0 ranges. Vowel: /ɑ/. A direct linear relationship is apparent. Each sunflower represents one or more cases that occur very close together on the scatterplot, with each case represented as a sunflower petal. Each sunflower is positioned at the mean of the cluster of cases that it represents. Each dimension of the sunflower cell is 1/15 of the appropriate range.

thought the pitch was "just normal," but listeners were told that they would probably need to use all the numbers equally often. They were able to press a "repeat" button to hear a stimulus again, but they were discouraged from listening more than twice. They were also allowed to rest at any point if they wished.

After the task had been completed, listeners were debriefed regarding the perceived difficulty of the task and strategies they believed they had employed. Additionally, all were asked to confirm that they had correctly understood the instruction to indicate where each token lay within its speaker-specific range. All of them did so.

III. RESULTS

A. Correlations by speaker tessitures

A linear regression was run on two variables, with assigned *ranking* of individual tokens (12 listeners) as the dependent variable and, as the independent variable, F0 *location* of the same tokens (calculated as a percentage of the relevant speaker-specific tessiture based on a log base 10 transformation of the high and low F0s from the glissandos). In the scatterplot given as Fig. 2, a relatively dense cluster (a binned range of values) of individual tokens appears as a collection of densely petalled "sunflowers" surrounding a regression line. Correlations were calculated for each listener separately. In all cases r was positive, ranging from 0.587 to 0.789, and $p<0.001$. In order to render the variance more independent of the population, Fisher's z transformation was applied to each r (Fisher, 1915). The mean of the resulting z values was calculated and that mean inverted into an r; this value was 0.721 and indicates a relationship between assigned rankings and location of the tones within speaker-specific tessitures.

B. Correlations by speaker tessitures, extrema excluded

In order to exclude the possibility that the correlation reported above was skewed by listener reliance on unusual voice qualities located at range extrema, we ran a further linear regression between assigned rankings from the middle six F0s for each speaker and actual locations of the middle six F0s within speaker-specific tessitures. That is, the high and low F0s for each talker were excluded. Half of these low values were judged by the experimenters to have a creaky voice quality. (Two other tokens that were second lowest were judged to have a creaky quality.) Here also, the scatterplot (see Fig. 3) suggests a linear relationship between ranking and location, and all r values were positive and significant ($p<0.001$); they ranged from 0.463 to 0.716. The inverse Fisher mean z (calculated as above) was 0.609, which indicates a direct relationship between the variables, though it suggests a weaker relationship than reported above for all eight F0s per speaker. However, a reduction in the strength of the correlation does not necessarily indicate that information around the center of each speaker's range was not sufficient for the listener. We return to this point in the Discussion.

C. Correlations by generic ranges

To test for the possibility that listeners were actually locating F0 relative to F0 ranges heard elsewhere (and perhaps separately for the two sexes), two further analyses were performed. One assumed that listeners would be able to determine for each stimulus whether a male or a female had uttered the vowel. Listeners are good at doing this based on sentence-length stretches of speech (Lass *et al.*, 1978), and on short segments of speech (Ingemann, 1968). It is also

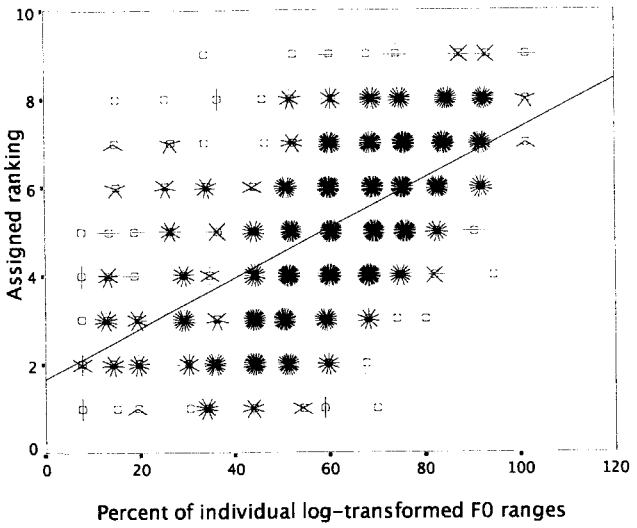

FIG. 3. Sunflower scatterplot with regression line between rankings assigned by 12 listeners and log base 10 transformed locations of six spoken tones per speaker (20 speakers) with locations reported as a percentage of speaker-specific spoken F0 ranges. The single highest and lowest tone for each speaker have been excluded from the regression, along with corresponding rankings. Vowel: /ɑ/. A direct linear relationship is apparent. Plotting dimensions are as in Fig. 2.

possible to reliably distinguish sex on the basis of the acoustics of vowel segments alone (Childers and Wu, 1991). All of these studies use central F0 values, so it is not clear whether the current listeners would be able to make this distinction throughout the range. Nonetheless, it seemed reasonable to assume that such a distinction might be of perceptual value, so a correlation was calculated as before but using the male or female extreme values as the definition of the range rather than the individual speaker's extremes.

The pattern of correlations is very similar to that found before, so no figure is given. The correlations for the sex-linked estimates were all positive and significant ($p<0.001$), ranging from 0.611 to 0.901, and the inverse Fisher mean z was again positive (0.830, $p<0.001$). The magnitude of the correlation was somewhat larger than that for the individual tessitures.

A final analysis used the overall extreme values (low from the lowest male, high from the highest female) as defining the global "tessiture." The correlation coefficients for the global, non-sex-linked range were also positive and significant ($p<0.001$), ranging from 0.611 to 0.934, with an inverse Fisher mean z of 0.848. These correlations are similar to those found in van Bezooyen (1984) for mean F0 and the perceived pitch level of her natural speech stimuli (expressing nine emotions plus a neutral reading).

D. Comparison of the two types of ranges

In order to assess the fit between the estimate of the location within range and the ranges based on individual and group tessitures, we have plotted the eight sustained F0s collected from each speaker, superimposing speaker-specific upper and lower tessiture boundaries elicited from spoken glissandos, plotting data according to the value of the bottom of the tessiture (Fig. 4). Colors in the figure represent a continuous gradation based on averaged listener rankings. The figure allows some comparison of success with the two types of range. Speaker-specific patterns of color updrift across the figure are balanced by global patterns seen in the relative stability of the mid-range bands of color across speakers.

Colored dots tend to expand to fill the tessitures whose boundaries are demarcated by black lines, though there are notable gaps between the lowest or highest F0 attained and the limit of the speaker-specific tessiture. That is, there are steady F0s that were produced and excluded either because eight values had already been obtained from that speaker or because no acceptable token was obtained by our elicitation technique. Those gaps are seen especially, on the low end, in the productions of speakers 6 and 8 among the males and 4, 11, 12, and 17 among the females, and on the high end, in the productions of speakers 2, 3, 6, 9, 18, and 20 among the males and 15 and 16 among the females. The lack of a dark blue dot for the lowest production of speakers 7, 9, 13, and 15 is more consistent with a global interpretation of F0 range fitting, since these attained F0s were close enough to the lower limit of the tessitures that listeners ought to have been able to assign them a ranking of 1 if they were able to locate F0 within individual ranges. Here, just as the correlation analyses show, there is evidence for both specific and general patterns.

IV. DISCUSSION

Significant positive correlations obtain between the assigned ranking and the actual location of tones within individually defined tessitures. However, similar correlations were also found when the F0s were scaled to the ranges of the speakers as a group. Since there was sizable overlap among the F0 ranges of our current talkers, it was not possible to distinguish which factor was more important.

One possible objection to the analysis might be that listeners were biased by the inclusion of extrema in the random presentation of F0s. Therefore, in order to rule out the possibility that ranking of F0 extrema was accurate but ranking of more central F0 values inaccurate, linear regressions on individually defined tessitures were rerun with the top and bottom F0 for each speaker excluded. This test produced slightly lower correlation coefficients (inverse Fisher mean z of 0.609, down from 0.721), but, crucially, the correlation was still positive and significant at $p<0.001$, even though n was 25% smaller. This test confirms the finding that listeners' successful ranking of pitch did not depend upon identifying peculiarities associated with the extreme high or low F0 values for each speaker.

The fact that listeners are able to locate within a tessiture steady, isolated pitches produced by unfamiliar voices implies some degree of speaker normalization. Although the present results strongly support our hypothesis that parsing of the F0 location does not require context or familiarization, they do not allow us to distinguish between possible models of F0 normalization. Listeners might build models of individual speakers or they might fit individual pitch ranges to previously built prototypes of possible vocal ranges. It is known that factors extrinsic to the signal are implicated in normalization-factors such as the overall formant range (Ladefoged and Broadbent, 1957) or the acoustics of inspi-

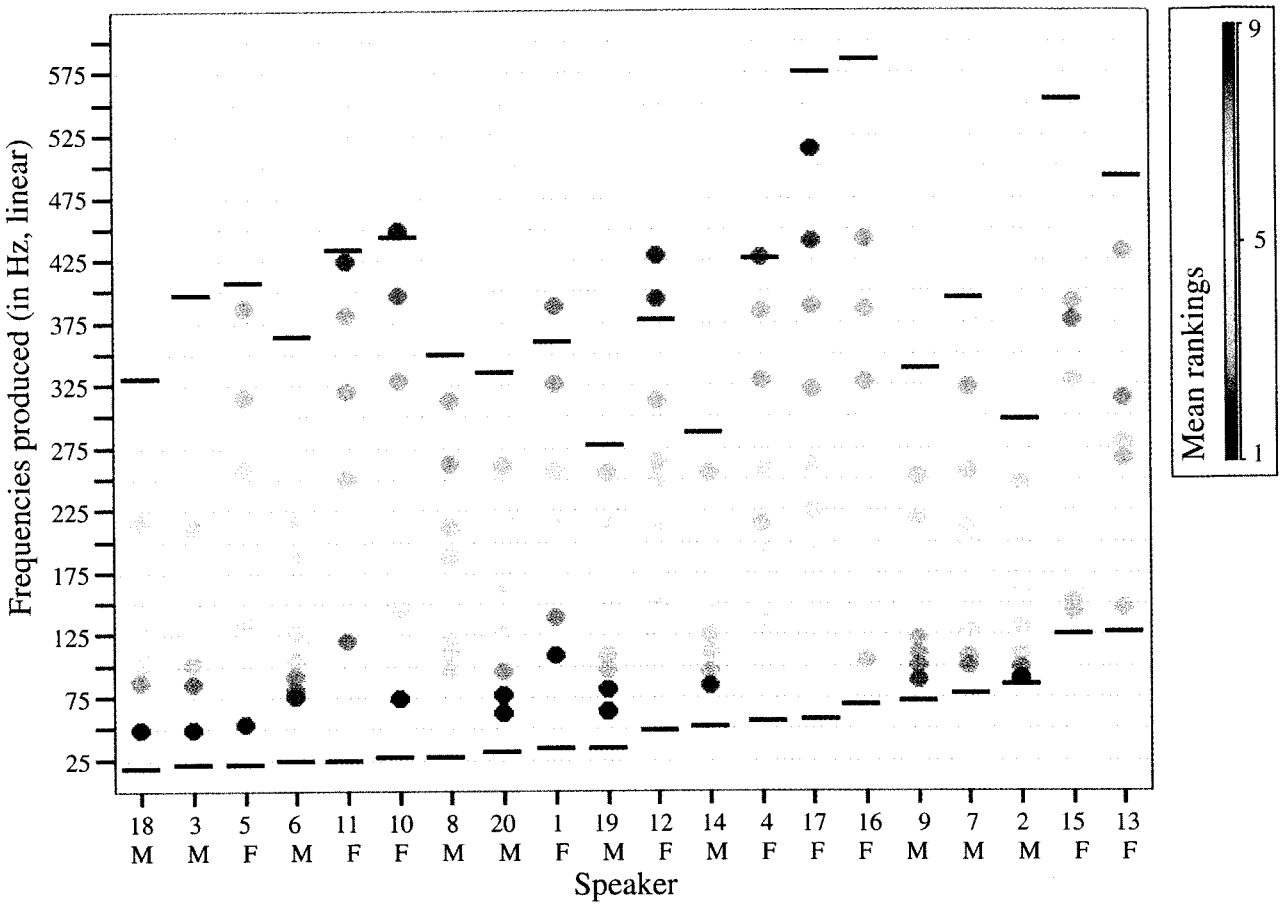

FIG. 4. Color-scaled mean rankings (12 listeners) plotted for each of eight sustained frequencies produced by each of 20 speakers. Data are ordered by bottom of glissando-based tessiture. The sex of the speaker is given below the speaker number.

ration, for example (Whalen and Sheffert, 1997). If signal-extrinsic aspects of speech are used in speaker normalization, it stands to reason that signal-intrinsic aspects of speech might also be used as listeners familiarize themselves with the voices of individual vocal tracts or vocal tract types. Voice quality is one such a signal-intrinsic feature.

Extremely low F0 within a speaker's range is often associated with a glottal fry or creaky voice (Ladefoged, 1971; Laver, 1980). Similarly, breathy voice tends to occur with low F0s (Fairbanks, 1960). Thus it may be that creak and breathiness indicate to the listener that the lower portion of the tessiture is being used. At the high end, the mode of vibration of the vocal folds changes dramatically between modal and falsetto voice (Hollien *et al.*, 1971; Laver, 1980). Minimally, therefore, it is likely that listeners can use information about voice quality directly to gauge the location of a steady pitch at the very least at the top and bottom of a speaker-specific range. However, in the present study, listeners appeared to base their rankings on signal-intrinsic (perhaps voice quality) information at or near the extrema as well as throughout the tessitures.

The proposal that listeners rely at least somewhat on voice quality as they locate pitch within a speaker-specific range does not lead naturally to a relational theory of F0 normalization in which pitch is located relative to other pitches via a Fraction-of-Range comparison or Z-Score-around-Mean comparison (e.g., Rose, 1987). In fact, one finding in the present investigation suggests the possibility of a more direct route to F0 normalization: Listeners performed well throughout a speaker's range, even though F0s were presented randomly. If listeners use information on voice quality intrinsic to the signal to aid in predicting pitch location within the range, no model of speakers is needed, personal or generic. If listeners parse voice quality information and relate F0 information to it, normalization may follow a more direct route. The present findings do not argue conclusively in favor of voice quality as the basis of parsing, but they show clearly that listeners need not depend on sampling the F0s of a speaker before making range judgments. The slight improvement in the correlations when the F0s were coded for the full range of F0s in the experiment suggests that listeners might be calibrating their responses to their experience with typical F0 ranges that they have heard throughout their lives, with or without reference to the intricacies of voice quality changes at different parts of the F0 range. Since our speakers seem to be typical in this regard, such a strategy seems to be a reasonable one. Further tests, in which speakers with less overlapping ranges are selected (if this is possible), would be necessary to sort this out.

In addition to voice quality, there are many sources of information about pitch range available to listeners. However listeners use that information, we can be fairly certain that they do not perform normalization simply by linking speaker pitch location to the perceived size of the speaker. While it appears that vocal tract length in humans depends on skull size, and while it has been shown that vocal tract length, in

turn, directly affects the averaged distance between successive formant frequencies when the cross-sectional area is held constant (Fant, 1960; Fitch, 1997), neither average nor basal pitch (e.g., Cooper and Yanagihara, 1971) correlates directly with adult human vocal tract size. The perhaps common but false supposition of a body size-pitch connection may follow from our more generally correct intuition that large objects resonate at low frequencies, combined with a casual confusion between source and filter properties. Indeed, certain species of birds and mammals "falsely advertise" their body size with disproportionately low calls produced primarily by enlarging the vocal tract (Fitch, 1999; Fitch and Reby, 2001), but actual correlations between body size and the length and mass of the vocal folds have been found to be small in adult humans (e.g., Lass and Brown, 1978; Künzel, 1989) and in other vertebrates more generally (see Fitch, 1997 for a review). In fact, the larynx develops independently of the head and body, and is even hypertrophied in post-pubescent human males (Hollien, 1960). It seems that there is no direct link between larynx size (and thus average or basal F0) and vocal tract length or body size, so pitch normalization cannot proceed from "on-line" (probabilistic) estimation of vocal tract length.

Although F0 does not directly correlate with vocal tract length (or formant distance), a speaker's sex can be determined reliably from either long (Lass *et al.*, 1978) or short (Ingemann, 1968) segments of typical speech, and automatic techniques are similarly successful with just vowels (Childers and Wu, 1991). It would appear that information regarding speaker sex may be present in what we might generally refer to as source or hybrid source-filter properties. It has been claimed, for example, that information about sex is available to the listener in the absolute value in Hz of some aspect of F0, usually mean or median F0. Peterson and Barney (1952) report male/female differences of around 100 Hz, as does van Bezooyen (1984). Indeed, Bachorowski and Owren (1999) found that sex was classified correctly 92.4% of the time by discriminant analysis on the basis of formant frequency cues to vocal tract length for 2500 tokens of one vowel produced by 125 subjects. The success of a multivariate statistical technique in classifying speaker sex on the basis of vocal tract length cues alone does not necessarily mean that listeners judge speaker sex in the same way, or that listeners use judgments of speaker sex to normalize pitch at a gross level, but the possibility remains. It has also been claimed that information about sex may be available in spectral tilt. For example, Hanson and Chuang (1999) found the ratio of the amplitude of H1 to that of F3 to be 9.6 dB lower for male than for female speakers, with lower mean values and less variation between male speakers for F1 bandwidth, open quotient, and other measures. However, it is not known whether information about speaker sex is available to the listener in short segments throughout the tessiture. It may be that F0s at the high or low end can be misleading about the sex of the talker, with low values perhaps sounding more typically male and high values, female. This question will be addressed in future research. For now, it is reasonable to assume that the sex of the speaker for at least some of the stimuli could have been perceived in this experiment.

Given the listener's potential for identifying the sex of the speaker, we might anticipate the following interpretation of our results, namely that listeners rank F0s based on stored templates representing population tessitures for adult male and adult female speakers, respectively. Indeed, correlations based on F0 locations within sex-based population ranges were almost as high as those based on global (sex-independent) ranges. Figure 4 shows evidence for both specific and global trends. If listeners were simply fitting judgments to absolute high- or low-frequency values, the colors of the dots would be expected to appear in straight horizontal bands uniformly throughout the chart; they do not, especially at the high and low ends. Yellow and near-yellow bands in the mid-tessiture range do cluster roughly in the 125–225 Hz range, but so do tessiture midpoints. In other words, where the yellow bands seem too high or too low relative to steady F0s produced by a speaker, they tend to fall roughly on center relative to speaker-specific tessitures. Furthermore, if listeners were simply fitting judgments to male and female (adult) population tessitures, the colors would be expected to band within sex, but plotting them this way (not shown directly) does not reveal any clear difference. In most cases where the lowest F0 within a speaker-specific tessiture does not receive a blue ranking or the highest F0, a red ranking, it is because of gaps in the stimuli. That is, there are steady F0s that, in an ideal world, could have been produced, but were not produced by some of our speakers. Major gaps appear on the low end for speakers 6 and 8 among the males and 4, 11, 12, and 17 among the females, and, on the high end, for speakers 2, 3, 6, 9, 18, and 20 among the males and 15 and 16 among the females. If the reader imagines dark blue and dark red dots in these spaces, respectively, a sex-independent pattern is enhanced. The dark dots simply follow tessiture boundary marks; they do not form bands of a single color family across all speakers of a particular sex. The lack of very high rankings may also be due to our intentional exclusion of falsetto from the stimuli. To the extent that falsetto is part of the speech range, listeners would be correct in not assigning a score of 9 to the highest F0s presented.

Where the F0 literature contains multiple studies of mean or median fundamental frequency as noted above, we focused on the entire range, and found a considerable overlap in tessiture between male and female speakers both in range and location of range. The overlap is especially apparent at the low end, where there is very little in the way of a consistent sex-based F0 difference in the glissando base. What we do see are more gaps for the females, who seem to have had a difficult time sustaining low F0s to which they can indeed glide down in a spoken voice register. At the high end, the highest glissando breaks are achieved by the females, and there is less overlap between sexes. There is some overlap, however, but it is again accompanied by (complementary) high-end gaps in sustained-but-possible F0s by the males. Had listeners here relied on a first-pass discrimination of sex to provide two templates for comparison, they would have misranked far more of these tokens. This pattern is consistent with the evidence for the use of both individual and global ranges.

Whether or not voice quality alone routinely provides

listeners with sufficient information for F0 normalization, it is clear that listeners are sensitive to the location of pitch within a speaker's range, even for isolated utterances. It will take further experimentation to isolate the features that listeners are sensitive to and to explain why listeners are good but not perfect at the task. For example, it would prove useful to test whether varying synthetic voice quality across a range of F0s confounds the listeners' attempts to track F0 values. Another avenue of research suggested by the present study would involve a replication in which the experimenters selected speakers with tessitures located high and low in absolute frequency and with less overlap than seen among the present speakers, if such speakers can be found.

ACKNOWLEDGMENTS

We thank Jeff Weihing for assistance with MATLAB programming and Brook Swainson and Paul De Decker for their assistance with stimulus extraction. We also thank Carol Fowler, Gordon Ramsay, and Yi Xu for their comments and Bob Ladd for a helpful discussion of the data reported here. During the review process, this paper also benefited from a number of insights offered by the Peter Assmann, Tecumseh Fitch and an anonymous reviewer. Any shortcomings in the present research remain our own. This work was supported by NIH Grant No. DC-02717 and NIH Grant No. DC-03782 to Haskins Laboratories and by an equipment loan from the Institute for Voice and Speech.

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

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

Cooper, M., and Yanagihara, N. (**1971**). "A study of the basal pitch level variations found in the normal speaking voices of males and females," J. Commun. Disord. **3**, 261–266.

Di Paolo, M., and Faber, A. (**1990**). "Phonation differences and the phonetic content of the tense-lax contrast in Utah English," Lang. Var. Change **2**, 155–204.

Fairbanks, G. (**1960**). *Voice and Articulation Drillbook* (Harper and Row, New York).

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

Fant, G., Liljencrants, J., and Lin, Q. (**1985**). "A four-parameter model of glottal flow," Speech Transmission Laboratory—Quarterly Progress and Status Report, Royal Institute of Technology, Stockholm 4, pp. 1–13.

Fellowes, J. M., Remez, R. E., and Rubin, P. E. (**1997**). "Perceiving the sex and identity of a talker without natural vocal timbre," Percept. Psychophys. **59**, 839–849.

Fisher, R. A. (**1915**). "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population," Biometrika **10**, 507–521.

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

Fitch, W. T. (**1999**). "Acoustic exaggeration of size in birds via tracheal elongation: Comparative and theoretical analyses," J. Zool. **248**, 31–48.

Fitch, W. T., and Reby, D. (**2001**). "The descended larynx is not uniquely human," Proc. R. Soc. London, Ser. B **268**, 1669–1675.

Gelfer, M. P. (**1989**). "Stability in phonational frequency range," J. Commun. Disord. **22**, 181–192.

Gobl, C., and Ní Chasaide, A. (**1992**). "Acoustic characteristics of voice quality," Speech Commun. **11**, 481–490.

Goldinger, S. D. (**1996**). "Words and voices: Episodic traces in spoken word identification and recognition memory," J. Exp. Psychol. Learn. Mem. Cogn. **22**, 1166–1183.

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.

Heylen, L., Wuyts, F. L., Mertens, F., De Bodt, M., and Van de Heyning, P. H. (**2002**). "Normative voice range profiles of male and female professional voice users," J. Voice **16**, 1–7.

Hollien, H. (**1960**). "Some laryngeal correlates of vocal pitch," J. Speech Hear. Res. **3**, 52–58.

Hollien, H., Brown, W. S., and Hollien, K. (**1971**). "Vocal fold length associated with modal, falsetto, and varying intensity phonation," Folia Phoniatr. **23**, 66–78.

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

Johnson, K., and Mullennix, J. W. (ed.). (**1997**). *Talker Variability in Speech Processing* (Academic, San Diego, CA).

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.

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

Ladefoged, P. (**1971**). *Elements of Acoustic Phonetics* (University of Chicago Press, Chicago).

Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 98–104.

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

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

Laver, J. (**1980**). *The Phonetic Description of Voice Quality* (Cambridge University Press, Cambridge).

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

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

Nygaard, L. C., and Pisoni, D. B. (**1998**). "Talker-specific learning in speech perception," Percept. Psychophys. **60**, 355–376.

Palmeri, T. J., Goldinger, S. D., and Pisoni, D. B. (**1993**). "Episodic encoding of voice attributes and recognition memory for spoken words," J. Exp. Psychol. Learn. Mem. Cogn. **19**, 309–328.

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

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

Remez, R. E., Fellowes, J. M., and Rubin, P. E. (**1997**). "Talker identification based on phonetic information," J. Exp. Psychol. Hum. Percept. Perform. **23**, 651–666.

Rose, P. (**1987**). "Considerations in the normalisation of the fundamental frequency of linguistic tone," Speech Commun. **6**, 343–351.

Schutte, H. K., and Seidner, W. (**1983**). "Recommendation by the Union of European Phoniatricians (UEP): Standardizing voice area measurement/phonetography," Folia Phoniatr. **35**, 286–288.

Sheffert, S. M., and Fowler, C. A. (**1995**). "The effects of voice and visible speaker change on memory for spoken words," J. Mem. Lang. **34**, 665–685.

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

Titze, I. (**1995**). "Speaking vowels versus singing vowels," J. Singing **52**, 41–42.

van Bezooyen, R. (**1984**). *Characteristics and Recognizability of Vocal Expressions of Emotion* (Foris Publications, Dordrecht).

Whalen, D. H., and Sheffert, S. M. (**1997**). "Normalization of vowels by breath sounds," in *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullennix (Academic Press, San Diego), pp. 133–144.

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.

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.