The direct and indirect roles of fundamental frequency in vowel perception

Santiago Barreda[a] and Terrance M. Nearey
Department of Linguistics, University of Alberta, Edmonton, Alberta T6G 2E7, Canada

(Received 30 December 2010; revised 18 May 2011; accepted 19 October 2011)

Several experiments have found that changing the intrinsic f0 of a vowel can have an effect on perceived vowel quality. It has been suggested that these shifts may occur because f0 is involved in the specification of vowel quality in the same way as the formant frequencies. Another possibility is that f0 affects vowel quality indirectly, by changing a listener's assumptions about characteristics of a speaker who is likely to have uttered the vowel. In the experiment outlined here, participants were asked to listen to vowels differing in terms of f0 and their formant frequencies and report vowel quality and the apparent speaker's gender and size on a trial-by-trial basis. The results presented here suggest that f0 affects vowel quality mainly indirectly via its effects on the apparent-speaker characteristics; however, f0 may also have some residual direct effects on vowel quality. Furthermore, the formant frequencies were also found to have significant indirect effects on vowel quality by way of their strong influence on the apparent speaker.

© *2012 Acoustical Society of America*. [DOI: 10.1121/1.3662068]

PACS number(s): 43.71.An, 43.71.Es, 43.71.Bp [AA] Pages: 466–477

I. INTRODUCTION

Listeners are able to recognize the vowels of their language with relative ease despite the fact that the physical characteristics of these sounds can vary a good deal from speaker to speaker. However, the same variation that hinders speech perception affords the listener with a wealth of information regarding the speaker. For example, listeners are able to judge the gender of an adult speaker with relative ease (Bachorowski and Owren, 1999; Strand, 2000; Perry *et al.*, 2001). They are also able to make consistent judgements regarding the apparent size of the speaker using only information available from that speaker's voice[1] (van Dommelen and Moxness, 1995; Lass *et al.*, 1980; Collins, 2000; Smith and Patterson, 2005; Smith *et al.*, 2005; Rendall *et al.*, 2007). If, and how, apparent-speaker characteristics and phonetic information interact in the determination of speech sounds is an open question in speech perception.

From the perspective of speech production, the fundamental frequency (f0) and formant frequencies (FFs) of a vowel are more or less independent (Fant, 1960) so that f0 should have only a small effect on the spectral content of a vowel. If vowel quality were entirely determined by the spectral content of a vowel, a change in f0 alone should cause no change in vowel quality. However, many experiments have induced vowel quality changes by changing intrinsic and/or extrinsic f0 with respect to the vowel's FFs (Miller, 1953; Fujisaki and Kawashima, 1968; Slawson, 1968; Johnson, 1990; Glidden and Assmann, 2004). Studies have induced similar effects by changing only the expected gender of the speaker. This has been done by presenting alternating male and female faces with identical stimuli (Glidden and Assmann, 2004) and by telling listeners to imagine either a male or female speaker (Johnson *et al.*, 1999). It is not clear why changes in apparent-speaker characteristics and changes in f0 affect vowel quality or if they do so independently or via the same general mechanism.

There are three general schools of thought regarding the relationship between vowel quality, f0 and apparent-speaker characteristics. These will be referred to as direct f0 theories, indirect f0 theories, and f0-free theories.

A. Direct f0 theories

Average f0 tends to co-vary with average FFs across speakers (Hollien, 1994; Fitch and Giedd, 1999; Nearey and Assmann, 2007). In general, larger people have lower FFs and f0s while smaller people have higher FFs and f0s. Listeners have been found to show a sensitivity to this covariance. They rate speech as more natural (Assmann and Nearey, 2007) and identify vowels correctly at a higher rate (Lehiste and Meltzer, 1972; Gottfried and Chew, 1986; Assmann and Nearey, 2008) when speech has the expected relationship between f0 and FFs. For these and related reasons, some researchers suggest that listeners take advantage of this covariance and, as a result, that f0 is directly related to vowel quality in the same way the FFs are (Syrdal and Gopal, 1986; Miller, 1989). These theories will be referred to as direct f0 theories. The net effect of these theories is empirically indistinguishable from one in which f0 is used by listeners as a scaling factor to eliminate inter-speaker differences by interpreting FFs in relation to f0 (Nearey, 1989, 1992).

B. Indirect F0 theories

Others, such as Nearey and Assmann (2007) and Johnson (1990, 1999, 2005), suggest that f0 is most important in determining certain apparent-speaker characteristics rather than in the specification of vowel quality directly. According to these theories, f0 is related to vowel quality only insofar

[a] Author to whom correspondence should be addressed. Electronic mail: sbarreda@ualberta.ca

as it contributes to the determination of whichever apparent-speaker characteristics affect vowel quality. These theories will be referred to as indirect f0 theories. Johnson (1990) suggests that listeners create a mental representation of the speaker and that speech is interpreted on the basis of the characteristics of this presumed speaker. In this model, f0 is only used to determine likely speaker identity. Johnson (2005) takes this several steps further and outlines an exemplar based "talker normalization" model:

"Rather than warp the input signal to match a fixed internal template, the internal representation adapts according to the 'perceived identity of the talker' (Johnson, 1990), as exemplars appropriate for the talker are activated and inappropriate exemplars are deactivated. […] cues of all kinds can be involved in tuning the activated set of exemplars [… including] F0 as a gender cue" (383).

Other researchers suggest that indirect normalization takes place via more abstract apparent-speaker characteristics rather than properties tied to limited classes of exemplars. For example, the probabilistic sliding template model (PSTM) (Nearey and Assmann, 2007) works on the basis of Ψ, which is a speaker-dependent value roughly equivalent to the average FF produced by a speaker. By adding Ψ^*, an estimate of Ψ, to a language-specific reference pattern, a listener can estimate expected FF for the vowels of that language as produced by a speaker. The PSTM uses f0, as well as information about the distribution of average FFs and the relationship between FFs and f0 to estimate the most likely Ψ for that speaker. [See also Traunmüller (1994) for a rather more elaborate account of an indirect relationship between observed f0 of a specific stimulus and perceived vowel quality; this approach may make predictions similar to those of the indirect normalization theories considered above, at least in some circumstances.]

C. f0-free theories

A final possibility is that there is no relationship between f0 and vowel quality. These theories will be referred to as f0-free theories. Despite the results of experiments reported in Sec. I B above, Patterson and colleagues have made strong claims about the independence of f0 and vowel quality. In a series of experiments that manipulate spectrum envelope and f0 independently via a vocoder, they found that changes in f0 have virtually no effect on vowel quality[2] (Smith *et al.*, 2005).

To explain this, Smith *et al.* (2005) and Irino and Patterson (2002) have suggested that the auditory system performs a Mellin(-like) transform on the acoustic input at an early stage in auditory processing. This results in a *size-shape image* (Irino and Patterson, 2002; 188) in which the spectral pattern of a sound is represented as an invariant shape and the size of the resonator that produced the sound is represented as the position along one dimension of the sound pattern in the sound-shape image. In this view, changes in f0 or in apparent-speaker properties play no role in determining vowel quality.

D. Rationale for the present study

All three of the above theories could be considered different forms of vowel normalization, where normalization refers to a process by which a listener removes or compensates for speaker-specific variation from an incoming vowel token. We are treating the normalization process as a black box where we may observe the input (the physical properties of the stimuli) and the output (vowel quality) but not the internal workings of the system. We do not seek here to determine the exact internal workings of the normalization process, but simply to consider what kinds of information may affect the transfer characteristics of the process.

The experiment to be described in the following pages was designed to test the relationship between f0, vowel quality and apparent-speaker characteristics. To do this, a vowel continuum was matched with several different f0s and higher formants (in this case, formants higher than F2 which will be referred to as F3+). The general stimulus design is similar to that of Fujisaki and Kawashima (1968) and to the isolation condition in experiments described in Johnson (1990). In fact, the experiment to be outlined here could be viewed as an extension and refinement of some of the experiments described in Johnson (1990). Because of the importance of some of the results presented in that paper to our current experiment, some of the relevant results will be summarized.

Johnson used a series of synthetic /hVd/ tokens with varying formant and f0 levels which were intended to be interpreted as either /ʌ/ or /ʊ/.Vowels were presented in two conditions: an isolation condition and a phrase condition. In the isolation condition, vowels were presented in a random order (with no extrinsic context) so that the intrinsic f0 of a vowel stimulus varied randomly from trial to trial. This would have resulted in something like a "speaker-randomized" condition. In the phrase condition, the same /hVd/ stimuli were presented following a synthetic voice saying "this is", which had either a rising intonation (simulating a question) or a falling intonation (simulating a declarative).

Johnson conducted an AX-discrimination pretest using stimuli with a single set of formant frequencies, but many f0 levels. Listeners were presented with pairs of stimuli and asked to judge whether the two syllables were spoken by the same or different speaker. Results indicated that although two tokens with the same f0 might be very likely to be judged as being from the same speaker in the isolation condition, the opposite is the case in the phrase condition since a speaker is unlikely to use the same final f0 for a phrase with falling and rising intonation. Johnson also conducted a second pretest, where listeners provided judgments of speaker size and gender for the stimuli of the AX pretest. The results provide evidence that size judgments are affected by the likelihood of perceived speaker differences as measured in the AX test.

Based on the results of the AX pretest, Johnson designed three vowel classification experiments involving a seven-member formant continuum and two f0 levels per experiment. These experiments were intended to test the relationship between apparent speaker changes and vowel perception. These experiments and the AX pretest were carried out with different groups of participants. Using this methodology, Johnson found an association between the likelihood of a perceived change in speaker in the pretest and the magnitude of an f0-induced vowel category shift in the main experiments. In listening conditions in which listeners were likely to hear different speakers, f0-induced shifts were maximized. In conditions in which listeners were likely to hear a single speaker these same effects were minimized. This association applied to both the isolated word and the phrasal presentation conditions.

Johnson presents a strong circumstantial case for the relationship between f0-induced vowel quality changes and apparent-speaker characteristics. Although his conclusions rely on some very reasonable inferences, they are inferences nonetheless. Specifically, the methodology does not allow for insight into the decisions listeners make on a trial-by-trial basis; nor, for that matter, does it allow for insight into the behavior of any one listener in both the pretests and the main experiments, since different listeners were involved in all cases.

The experiment to be described below represents, in a sense, an amalgamation of aspects of both pretests and of the isolation conditions of experiments 1 and 2 of Johnson (1990). For each stimulus presented, we asked participants to make simultaneous judgments of vowel quality and two aspects of speaker characteristics, so that analysis could proceed on a token-by-token basis. Johnson found large effects of f0 for isolated syllables in experiment 1, where f0 and formant patterns varied from trial to trial. When more information is available about an apparent speaker's intonation and (possibly) formant ranges, the effect of f0 on vowel quality may be greatly reduced.[3] Our experiment uses isolated vowels with complete randomization of all stimulus properties from trial to trial, resulting in what amounts to a speaker-randomized condition with little to no extrinsic context.

By simultaneously collecting both vowel quality information and apparent-speaker characteristics, we can relate f0-induced vowel quality shifts to changes in the apparent speaker. Although we are not asking for listeners to identify speaker changes directly, the collection of speaker gender and size information will allow us to control for important aspects of perceived speaker changes from the perspective of the listener at the moment of the vowel judgment.

If f0 and apparent-speaker characteristics do not contribute to the determination of vowel quality, they should not have a significant relationship to vowel quality after the formant frequencies have been accounted for. If f0 is directly related to vowel quality, there should be a stable and consistent relationship between f0, the FFs, and vowel quality. Additionally, after these physical properties have been taken into account, there should be no relationship between vowel quality and apparent-speaker characteristics. If f0 affects vowel quality mainly indirectly via its effect on apparent-speaker characteristics, there could be a variable and complicated relationship among judgments of apparent-speaker characteristics, f0 and vowel quality. Furthermore, the relationship between f0 and vowel quality should be considerably weaker, or perhaps non-existent, once apparent-speaker characteristics are controlled for.

II. METHODOLOGY

A. Participants

Listeners were 19 students from the University of Alberta, 16 females and 3 males drawn from a participant pool in which undergraduate students take part in experiments in exchange for partial course credit. They ranged in age from 17 to 54 years old. All were students taking an introductory level, undergraduate linguistics course.

B. Stimuli

The vowel continuum was constructed on the basis of naturally produced data collected from Edmonton English speakers. A continuum was designed that spanned from roughly the average F1-F2 frequencies of the /ʌ/ of a male to those of the average /æ/ produced by a female in seven equal logarithmic steps. The vowels used were chosen because, when produced by Western Canadian English speakers, they fall on a line almost exactly parallel to the line $F1 = F2$ in log-formant space. This meant that a single scale factor could be applied to both formants to either change vowel identity or to approximate the change in FFs because of a change in speaker size. Additionally, production data collected at the Alberta Phonetics Laboratory indicated that F3 was nearly identical for the two vowels, meaning that it carried little to no phonetic information. As a result of this F3 could be manipulated without greatly affecting the phonetic quality of the vowels, at least for vowel stimuli consistent with those of a single speaker. The low F3 level was set using perceptual data also collected at the Alberta Phonetics lab. An F3 frequency was selected at which the /ʌ/-/æ/ boundary was perpendicular to the $F1 = F2$ line so that F1 and F2 would contribute about equally to possible category boundary shifts.

The fourth point of this continuum had F1-F2 frequencies appropriate for either an /æ/ produced by an adult male or an /ʌ/ produced by an adult female. This seven-step continuum was combined with three different F3+ conditions and three different f0 conditions for a total of 63 different vowels. The stimuli were designed in a log space using ln (Hz) (the natural logarithm of the frequency in Hz). The frequencies of all of the continuum points, f0 and F3 levels used are presented in Table I.

1. F1 and F2 values

Since the vowels fall almost exactly parallel to the $F1 = F2$ line in log space, F1 and F2 were modified at the same rate and are therefore perfectly correlated. For this reason they will be treated as one variable, which for the sake of brevity will simply be referred to as F1. The formants for

TABLE I. Formant frequencies and f0s (Hz) used in the creation of the stimuli.

	f0 levels				F3 levels		
	Low	Mid.	High		Low	Mid.	High
Initial	120	170	240		2475	2755	3068
Final	96	136	190				
Step #	1	2	3	4	5	6	7
F1	684	735	789	848	911	978	1051
F2	1354	1455	1563	1679	1803	1937	2081

each successive step were about 0.0713 natural log units higher than those of its predecessor. This corresponds to an increase of about 7.4% in Hz. A three step difference in the F1 continuum corresponds to a 0.214 ln (Hz) change (a 22.5% increase). This is about one third the difference between the typical male /ʌ/ (step 1) and the typical male /æ/ (step 4) and also about one third the difference between the typical male /æ/ (step 4) and the typical female /æ/ (step 7). Therefore, a change of F1 of this magnitude [0.214 ln (Hz), or 22.5%] corresponds to the distance between these phonemes for a single speaker and to the average difference between the phonemes as produced by males and females.

2. *F3 and higher formants*

The low F3 was set at a value typical for adult males and the highest value was calculated by increasing the log frequency by the previously mentioned male to female step [0.214 ln (Hz), or 22.5%]. The intermediate F3 value is the (geometric) mean of the high and low F3s. The low F4 was set at 3200 Hz, and every successive FF (F5–F11) was set at 1100 Hz higher than the previous FF. The intermediate higher formant frequencies were raised by 11% relative to low higher formants, and high higher formants were raised by an additional 11% relative to the intermediate higher formant frequencies. The factor corresponding to F3 and the higher formants will be called F3+.

3. *Fundamental frequency*

The low f0 level was set to 120 Hz, appropriate for an adult male. The high f0 level was set to reflect the natural covariance between FFs and f0. Nearey and Assmann (2007) report that in a log scale, f0 increases 0.31 times as fast as typical FFs, which is close to the value of 1/3 used by Miller (1989) to relate the logs of F1 and f0. This means that, for example, a speaker who produces an average f0 1.0 ln (Hz) higher than a second speaker would also be expected to produce FFs that are 0.31 ln (Hz) higher (roughly 36%), on average, than this second person. In accordance with this relationship, the high f0 condition was one octave[4] higher than the low condition, which we set at a value appropriate for a male speaker. This resulted in a high f0 value of 240 Hz, which was considered appropriate for an adult female. The intermediate f0 condition is the (geometric) mean of the high and low f0s. The f0 values described above refer to the initial f0. The f0 contour decreased linearly across the vowel to a value 0.80 times the initial value.

The f0 levels in this experiment reflect the range observed for adults in Hillenbrand *et al.* (1995). Specifically, the lowest f0 used was 120 Hz which is about 0.51 standard deviations lower than the average male value (mean = 131 Hz, s.d. = 22 Hz). The highest f0 was about 0.84 standard deviations above the average adult female values observed (mean = 220 Hz, s.d. = 23 Hz).

4. *Synthesis of stimuli*

All vowels were 225 ms in duration, with steady-state formants. They were synthesized using an implementation of a Klatt (1990) synthesizer provided on version 5.1.01 of Praat (Boersma and Weenink, 2009) and synthesized at a sampling rate of 44.1 kHz. The Praat Klatt synthesizer works on the basis of tiers, each of which contains a separate piece of information about the sound to be synthesized. A single voice source tier was created containing the source specifications to be used for all vowels across all conditions. The source was created with a special focus on the female voice it would create, so that it would sound like a naturally produced female voice and not a male voice with a high pitch. This was accomplished by using a slightly breathy voice source and small negative spectral tilt, both of which have been found to be associated with femininity in North American English (Price, 1989; Klatt and Klatt, 1990; Mendoza *et al.*, 1996; Van Borsel *et al.*, 2009).

Three pitch tiers were created, one for each of the three f0 conditions. Tiers were also created containing formant frequency and bandwidth information for the higher formants, formants 3–11, in each of the three F3+ conditions. Because of the high sampling rate, 11 formants were found necessary to fill the Nyquist band and prevent excessive energy roll-off at higher frequencies. Formant bandwidths were set to the larger of 6% of the formant frequency or 60 Hz. All sounds were synthesized using the single voice source and every combination of formant and pitch tier for all three conditions, resulting in 9 distinct conditions (3 pitch conditions × 3 formant conditions).

C. Procedure

Participants were instructed that they would be hearing a human-like, “robotic” voice producing vowels intended to be either /ʌ/ or /æ/. Participants were asked to listen to the vowel and decide which of the two vowel categories the vowel sounded most like. In a pilot experiment, we asked participants to indicate how tall and how masculine/feminine the speaker they just heard was. We found that masculinity and femininity correlated strongly with f0 and that it may have been too specific a quality. Furthermore, many participants had difficulty reporting the height of the speaker. Some were not familiar with the imperial system (we asked for heights in feet and inches) while others felt that height was too specific; they thought the synthetic speakers varied by being more or less muscular or bulky rather than by being taller or shorter. Rather than ask participants for the continuous judgments of masculinity/femininity and height of the speaker, we asked participants for two kinds of judgments about apparent-speaker characteristics:

(1) A discrete gender judgement.
(2) A graded size judgment; the specific definition of size was left for the participants to interpret as they saw fit. The size judgement was intended to correlate with the listener’s estimate of the speaker’s vocal tract length, and hence formant ranges.

We left the definition of size deliberately vague because of difficulties encountered in pilot experiments that used absolute physical units. The lack of explicit instructions given to participants and the fact that the size scale might have been used in different ways within each gender may have led to differences in how listeners used the size scale (see the Appendix). However, any resultant increase in variability would only add noise to the data. It thus seems unlikely to bias any patterns in the data in any specific direction relevant to the hypotheses at hand.

Participants were presented with the sounds over headphones in a sound-attenuated booth and responses were recorded on a computer interface using software specifically designed by the first author for this experiment. Vowel quality responses were input by recording clicks of a mouse on a response button 800 pixels in length, where the *x*-axis coordinate of the pixel on which the participant clicked was entered as the response so that responses were recorded on an 800 point rating scale. Vowel responses were recorded on a button that said *Hud* (corresponding to /ʌ/) on one end and *Had* (corresponding to /æ/) on the other end. Participants were told that the selection of vowel had to fall into one category or the other and that clicking towards the extremes indicated the degree to which the vowel they had just heard sounded more like one vowel than the other. This scale was aligned so that a larger value corresponded to a more /æ/-like vowel. For this reason, this measure will be referred to as the openness of the vowel.

Speaker size responses were recorded on two separate buttons, one indicating a male speaker and one indicating a female speaker. Participants were instructed that selection of speaker size was also continuous and that clicking higher on the size button indicated a larger speaker. The size/gender buttons were 400 pixels high; in this case the *y*-axis coordinate at which the participant clicked was entered as the size response. The speaker-size judgment scale was aligned so that a larger value corresponded to a larger speaker. Size responses were recorded on two separate buttons, one labeled "male" and the other "female," which were placed orthogonally to the vowel response button. The use of two separate size buttons, one for each gender, allowed us to collect simultaneous gender and size information with a single click. Speaker gender was coded so that a value of 0 corresponded to a female speaker and 1 corresponded to a male speaker. Since this value indicates a male speaker, this value will be referred to as maleness. A screenshot of the experimental interface is provided in Fig. 1. To control for any spurious correlation between vowel and speaker judgments due to horizontal arrangement of the response buttons, the left-right position of the male and female response boxes was counter-balanced across listeners.

The procedure was as follows: A stimulus was presented, after which participants had to make a vowel quality judgment and indicate speaker size and gender. After these three values had been provided, the next stimulus would play after a 500 ms pause. Vowel sounds were presented in a random order along all stimulus dimensions. Participants were told they could repeat a stimulus up to 2 more times by hitting a button marked "replay" but only if they had not selected any responses for that stimulus. To cancel or undo any selections they had made, participants could click on a button marked "cancel" which erased any answers already provided for the current and previous stimuli, placed them both back into the upcoming stimuli queue, and re-shuffled the queue.

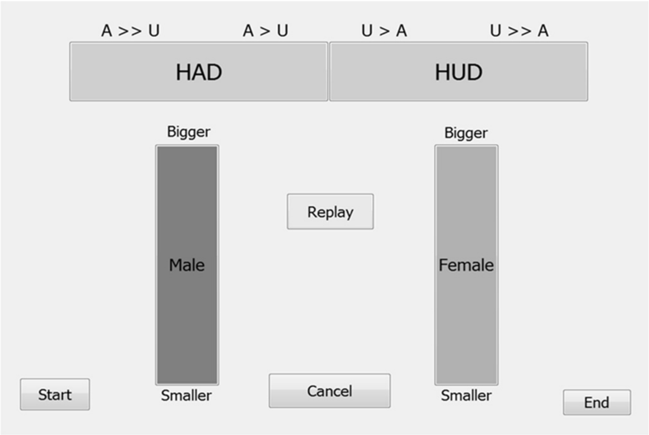

FIG. 1. Screenshot of the experimental interface.

Participants took part in experimental sessions of approximately one hour in length. Before beginning the experiment, participants completed a short training session during which they became familiar with the tasks and the response interface. During the training session participants heard naturally produced /hVd/ syllables containing either /æ/ or /ʌ/ in which the stimuli were produced by two male and two female speakers. Standard practice was to have participants listen to three repetitions of the stimulus list (189 responses), followed by a short break, after which the participant performed another three repetitions of the same list. In some cases, participants were not able to perform all six repetition of the stimuli list. In these cases, only the data from completed repetitions was used. A total of 6921 responses were collected across all 19 participants.

III. RESULTS

To organize a discussion of the results, we will outline the expected relationships between pairs of variables according to an indirect f0 theory in which f0 changes vowel quality by affecting a listener's frame of reference,[5] which in turn is assumed to be correlated with vocal tract length and formant frequency ranges. These relationships correspond to the expected correlations with *all other things being equal*. Open vowels occur with F1 frequencies near a speaker's maximum F1. A speaker with larger vocal tract has a lower maximum F1 than a speaker with a shorter vocal tract. If interpreted as coming from a speaker with a larger vocal tract, a stimulus with an intermediate F1 will appear to be nearer to that speaker's maximum F1 and hence sound more open. As a result, evidence which would lead a listener to conclude that the speaker is larger should lead to the perception of a relatively more open vowel, while evidence to the contrary would result in the perception of a relatively less open vowel for any given set of formant frequencies. A summary of related predictions is presented in Table II, assuming average natural relations between gender, f0 and vocal tract length.

This experiment contained three manipulated variables (F1, F3+, and f0) and three response variables (vowel openness, maleness, and speaker size). The manipulated variables were controlled experimentally and are not affected by any

TABLE II. Expected relationships between pairs of variables, all other things being equal. Where appropriate, the intermediate inference leading to this relationship is given.

Evidence	Inference	Expected Effect on variable
Higher F1		More open vowel
Higher F3+	Shorter vocal tract	Less open vowel
Higher f0	Shorter vocal tract	Less open vowel
Higher formants/f0	Less likely to be male	Female response
Higher formants/f0	Smaller speaker	Lower speaker size response
Larger speaker size	Longer vocal tract	More open vowel
Male	Longer vocal tract	More open vowel

other variables. The response variables are the three variables whose values are provided by the listeners. These reflect properties that exist only in the mind of the listener and may interact with the manipulated variables, and with each other, in unknown ways.

A. Partial correlation analysis

To investigate the relationship between these variables, a series of within-participant partial correlations was conducted. By considering the partial correlations between pairs of variables after controlling for all of the remaining variables, we can investigate the relationship between these variables independently (of any linear effects) of all the others. For example, the partial correlation between f0 and vowel quality after controlling for F1, F3+, speaker size and maleness will tell us how f0 and vowel quality are expected to co-vary for a vowel with given formant frequencies when produced by a speaker of given apparent size and gender. The process to be outlined below was carried out for each pair of response variable (vowel openness, maleness and speaker size) and every combination of individual response variable and individual manipulated variable (F1, F3+ and f0). The process will be outlined using the relation between f0 and speaker size as an example.

The following procedure was applied to the data of each listener in turn. To investigate the relationship between f0 and speaker size independently of all of the other variables in the experiment, each of these two variables was regressed in turn on the remaining four variables (F1, F3+, vowel openness, maleness). After this, the correlation between the residuals from the two regressions was found. The resulting partial correlation coefficient corresponds to the correlation between f0 and speaker size after controlling for the effects of all of the remaining variables. In this particular case, it is expected that f0 will be negatively related to speaker size since higher f0s should be associated with smaller speakers. If, all other things being equal, participants associate higher f0s with smaller speakers, then the partial correlation between speaker size and f0 should, on average, be significantly different from zero. If participants do not associate smaller speakers with higher f0s then the expected value of the average partial correlation between f0 and speaker size (after controlling for F1, F3+, and vowel openness) will be zero. Since this correlation is bi-directional, any discussion of cause and effect is dependent on the variables involved. For example, it is presumed that f0 causes the change in vowel openness rather than the other way around, since f0 is controlled by the stimulus design. Causal relations between pairs of judged qualities, however, are indeterminate.

This process was repeated for all 12 pairs of variables considered. This resulted in 19 partial correlation coefficients (one for each listener) for each of the 12 variable pairs. Following the two-stage procedure of Lorch and Myers (1990), independent sample t-tests were performed on the coefficients for every pair of variables to see if the results were significantly different from zero, on average across participants. The results of the t-tests are presented in Table III.

All except the last column of Table III relate directly to patterns predicted by the general indirect-f0 normalization model discussed at the beginning of Sec. III as summarized in Table II. Notably, all are in the expected direction and all are significant at $p < 0.01$ level or better, save for the relationship between speaker size and vowel openness.[6] Although the relation between speaker size and vowel openness does not reach significance using a t-test, 14 out of 19 speakers show a positive relationship between the two variables, a result that is not likely to have occurred by chance ($p = 0.022$ via a non-parametric binomial test).

The predictions of Table II involve relationships between specific stimulus properties and listener judgments of vowel quality or speaker characteristics, or between speaker characteristics and vowel quality. However, the last column of Table III involves the relation between the judgments of the two apparent-speaker characteristics, controlling for all other factors. The significant negative partial correlation between speaker size and maleness is at first surprising, since one would expect voices heard as male to be associated with larger absolute sizes. There are, it turns out, reasonable explanations for the negative partial correlation actually observed. These are discussed in the Appendix.

Figure 2 shows the distributions of the coefficients of Table III across listeners. In the discussion below, references to relative strength of the relationships between variable pairs will be based on the average magnitude (absolute value) of the partial correlation coefficient so that a variable pair with a larger magnitude will be deemed to have a stronger relationship than one with a smaller magnitude.

TABLE III. Results of t-tests performed on the within-participant partial correlation coefficients for pairs of variables. Variables included are F1, F3+, f0, vowel openness (V), maleness (M), and speaker size (S).

Relation	F1, VO	F3+, VO	f0, VO	F1, M	F3+, M	f0, M	F1, S	F3+, S	f0, S	M, VO	S, VO	S, M
Mean Corr.	0.802	−0.215	−0.053	−0.152	−0.147	−0.744	−0.212	−0.151	−0.374	0.049	0.027	−0.475
t (d.f. 18)	64.5	−10.9	−3.02	−6.49	−9.02	−41.0	−9.18	−4.05	−8.59	3.00	1.06	−12.2
p value	< 0.001	< 0.001	0.007	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.008	0.303	< 0.001

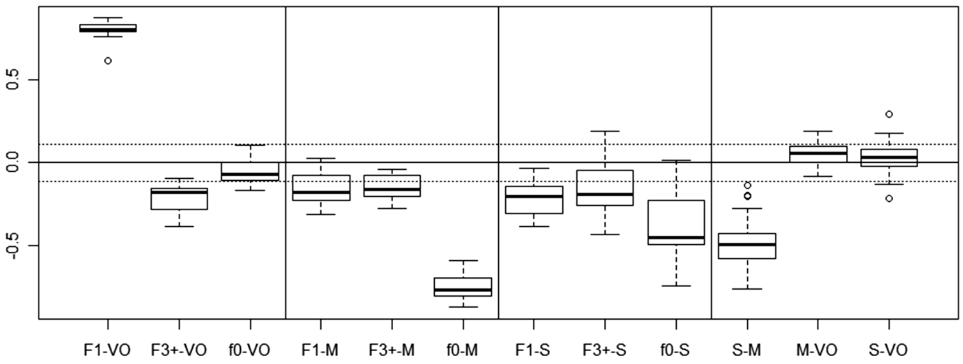

FIG. 2. Distributions (across participants) of average partial correlation coefficients between pairs of variables (VO = vowel openness, S = speaker size, M = maleness). The dotted lines represent bounds at which an individual participant's coefficient reaches significance ($p < 0.05$).

F1 and F3+ both relate strongly to vowel openness, though F1 is a stronger determinant. With the exception of one listener, the distribution of coefficients for the F1 to vowel openness relationship are tightly clustered around the mean, while the coefficients representing the relation between vowel openness and F3+ are more equally distributed over a wider area. The relations between F1 and maleness and F1 and speaker size are only slightly stronger than those between speaker size and F3+ and maleness and F3+. It seems that both F1 and F3+ affect both vowel quality and apparent speaker size, but that F1 is more strongly linked to vowel quality while F3+ is more strongly linked to apparent-speaker characteristics. Maleness is related to all three of the manipulated cues, though f0 is its strongest determinant. Speaker size is also determined jointly by considering all three manipulated variables and f0 is also its strongest determinant.

IV. ASSESSMENT OF THE DIRECTNESS OF EFFECTS

The fact that all of the relations presented in Table III are in the expected direction (all but one significantly so) is taken as evidence that the basic structure of the design was successful. Since the stimuli were synthesized using parametric synthesis, no real speaker identity or vowel quality can be associated with any of the stimuli other than whatever properties are attributed to the sound or speaker on the part of the listener. However, participants demonstrated an ability to extract both vowel quality and apparent-speaker characteristics from the stimuli. Furthermore, they interpreted this information in a fairly consistent way.

Figure 3 presents the same information found in Table III and Fig. 2 but in a manner that is easier to inspect visually. The arrows between variables indicate the presumed direction of the effects, and the numbers beside each variable indicate the average strength of the effects. The direct effect of a manipulated variable on the response variables can be judged by the average strength of the direct connection between the two variables. The indirect effect of a manipulated variable can be gauged by considering the effects the variable had on one or more of the response variables jointly with the effects the response variables have on each other.

Let us define a pure direct relationship between f0 and vowel openness as one that is not mediated by apparent-speaker characteristics. For example, the relationship between f0 and vowel quality in the model of Syrdal and Gopal (1986) qualifies as a pure direct relationship in this sense. The inclusion of concomitant information about a listener's impression of apparent-speaker characteristics should not affect this direct relationship in any way. Specifically, the correlation between vowel quality and f0 would be essentially unaffected after controlling for a listener's judgment of speaker gender in a partial correlation analysis.

Similarly we define a pure indirect relationship between f0 and vowel openness as one that is mediated by the direct effects of f0 on certain apparent-speaker characteristics: f0 affects the apparent-speaker characteristics which in turn affect vowel openness. In such a case, when behavioral measures of those apparent-speaker characteristics are accounted for, the partial correlation between f0 and vowel openness will approach zero.

The cases outlined above represent the endpoints of a range of possibilities. A series of exploratory models were considered which were intended to shed light on the relative direct and indirect effects of the three manipulated variables in the experiment (f0, F1, and F3+) on vowel openness.

To this end, we examined changes in partial correlation coefficients between manipulated variables and vowel openness in two kinds of models. We will illustrate these kinds of model for f0. The first kind of model will be referred to as a fully controlled model. It is identical in form to the kind of analysis reported in Sec. III A. To review, the partial correlation between f0 and vowel openness is calculated after controlling for all other variables; namely the two other manipulated variables, F1 and F3+, as well as the two other response variables, maleness and speaker size. The second kind of model

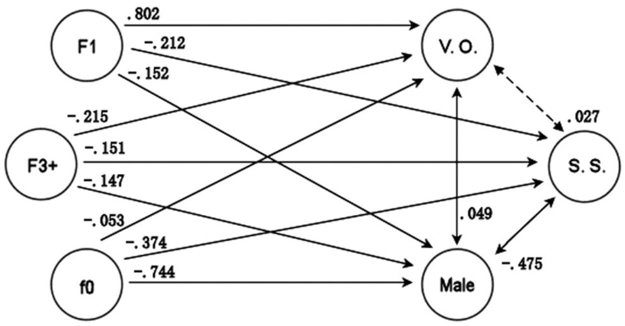

FIG. 3. Partial correlation coefficients (averaged over participants) between pairs of variables (V.O. = vowel openness, S.S. = speaker size, Male = maleness). The broken line between size and vowel openness indicates the only relationship which did not reach significance by t-test. Arrows indicate the presumed direction of effects.

will be called the no-speaker model, where the response variables maleness and speaker-size are left out of the model.

Thus the original, fully controlled model correlations included apparent-speaker characteristics, while the no-speaker model ignores them. Our analysis follows the logic outlined in the beginning of this section. If f0 has a largely direct relation to vowel openness, then there should be little difference in the partial correlations between f0 and openness of the fully controlled and no-speaker models. If the relation is predominantly indirect, then it is expected that the partial correlation coefficients between f0 and vowel openness will decrease noticeably in magnitude in the fully controlled model. The degree of this decrease will be taken as a measure of the relative indirectness of the relationship.

Similar assessments of the relative indirectness of the other two manipulated variables, F1 and F3+, were undertaken. For assessing the relation between F1 and vowel openness, f0 and F3+ were partialed out as control variables; for assessing the relation between F3+ and vowel openness, F1 and f0 were the control variables.

Differences between the two models will be tested using the same process outlined in Sec. III A, following the two-stage analysis of Lorch and Myers (1990). If the partial correlation coefficients do not change significantly between the two models, the expected value of the differences between the two estimated partial correlation coefficients for a single participant will approach zero. To test this, a series of paired *t*-tests were carried out on the differences between the two estimated coefficients across the 19 participants. The results of these *t*-tests show that all three differences are significant, indicating that the inclusion of apparent-speaker characteristics in the model significantly affects the relationship between vowel openness and F1, F3+, and f0. Furthermore, in all three cases the partial correlation coefficients as estimated by the fully controlled model decrease in magnitude relative to those obtained from the no-speaker model, indicating that F1, F3+, and f0 all have significant indirect effects on vowel quality. Of the three cues investigated, f0 was most strongly affected by the inclusion of apparent-speaker characteristics (mean difference $= 0.091$, $t = 4.48$, $\text{df} = 18$, $p\text{-value} < 0.0003$), followed by F1 (mean difference $= 0.021$, $t = 4.71$, $\text{df} = 18$, $p\text{-value} < 0.0002$) and F3+ (mean difference $= 0.017$, $t = 2.573$, $\text{df} = 18$, $p\text{-value} < 0.02$).

Another way to consider changes in the estimated coefficients across the two models is to consider the change in the mean partial correlation coefficient for pairs of variables between the no-speaker and fully controlled models. The means for pairs of variables across both models, and the corresponding percentage decreases in magnitude are presented in Table IV. Although the absolute change in the F3+ coefficient is smaller than that seen in the F1 coefficients, when this is considered as a percentage of its original magnitude, the relative change in F3+ is actually larger than that of the F1 coefficients. The change in the f0 coefficients is dramatically larger than either the F1 or F3+ changes. These results reinforce those presented in Sec. III A, which suggested that F1 was more strongly related to vowel openness than F3 and that f0 is strongly related to apparent-speaker characteristics but only a weak direct determiner of vowel openness.

TABLE IV. Mean partial correlation coefficients across all 19 participants for the fully controlled and no-speaker models. The percent decrease in mean indicates the decrease in magnitude from the fully controlled model to the no-speaker model relative to the magnitude of the no-speaker model.

	F1	F3+	f0
No-speaker mean	0.824	−0.232	−0.144
Fully-controlled mean	0.802	−0.215	−0.052
Decrease in magnitude	2.6%	7.5%	63.3%

V. GENERAL DISCUSSION

Since this experiment was designed to investigate the relationship between f0 and vowel quality, the first question is whether f0 affects vowel quality at all. It is clear that it does; participants identified an average of 11% more vowels as /ʌ/ when they had the highest f0 relative to the same vowels when presented with the lowest f0. This result is quite far from zero ($t = 6.1254$, $\text{df} = 18$, $p\text{-value} =< 0.0001$), and only 1 of 19 listeners did not show an increase in the number of vowels identified as /ʌ/ as f0 rose. The change in f0 must be ultimately responsible for the change in vowel quality across f0 levels since the vowels across f0 levels are identical in all other respects.

Not only does f0 have an effect on perceived vowel quality, but both sets of partial correlations considered in the previous section show a significant relationship between f0 and vowel quality after adjusting for other factors considered in either model. These results are difficult to reconcile with any hypothesis in which f0 is completely uncorrelated with vowel quality. Smith *et al*. (2005) and Irino and Patterson (2002) have proposed that vowel quality is entirely determined by aspects of the spectrum independent of f0. Since the partial correlation between f0 and vowel openness was calculated after correcting for F1 and F3 information,[7] and these factors should entirely determine vowel quality, it is not clear why f0 should have such a persistent relationship with vowel quality. In fact, our results indicate that any theory of vowel perception which completely disregards the influence of f0 on vowel quality cannot be an accurate representation of human behavior, at least in these random-speaker listening conditions.

The question then becomes whether the effect of f0 on vowel quality is mainly direct (as is the effect of the FFs) or mainly indirect (as is the effect of apparent-speaker characteristics).

If the effect of f0 on vowel quality were direct and based on the natural covariance of FFs and f0s experienced by people on a daily basis, then the relationship between these two variables, all other things being equal, should cluster around the value dictated by this natural covariance; it should not be spread over a large range of values. Additionally, the relationship between f0 and vowel openness should not be dramatically affected by controlling for relevant apparent-speaker characteristics. Specifically, if the relationship of f0 to vowel quality is of the same kind as the relationship between vowel quality and the formants, then the f0-vowel openness relationship and the F1-vowel openness and F3-vowel openness relationships should change in similar ways as a result of controlling for speaker size and gender.

Our results indicate that none of the restrictions or predictions posited by a direct f0 hypothesis play out. Participants show a wide range of sensitivities to this relation, in some cases even showing exactly the opposite relation between f0 and vowel quality than one would expect. Although the behavior of a few participants is unusual or difficult to interpret, the variation exhibited is itself a challenge to any theory of vowel perception in which f0 is tied to vowel quality in a stable and consistent way. If the effect of f0 is not fixed, but is instead modifiable to suit the listening conditions, then it ceases to be direct f0 normalization. This will also apply to any scheme that relies on fixed F1-f0 relations in the determination of vowel quality (see also Johnson, 1990). Furthermore, the relationship between f0 and vowel openness is considerably weakened after controlling for apparent-speaker characteristics while the F1-vowel openness and F3-vowel openness relationships maintain much of their strength. Although this does not tell us about the exact relationship between f0 and vowel openness, it is enough to conclude that this relationship is of a different kind than that between the FFs and vowel openness.

The hypothesis that f0 affects vowel quality mainly indirectly, via its effect on apparent-speaker characteristics is perhaps the only remaining viable hypothesis, and its predictions are well-supported by our results. Although f0 strongly affects vowel quality, once apparent-speaker characteristics have been accounted for (using the response variables speaker size and maleness) the relationship between f0 and vowel quality is weakened. Additionally, both speaker size and maleness show a consistent relationship with vowel openness independently of the FFs and f0. It seems that f0 affects vowel quality insofar as it affects a listener's expectations about the presumed speaker. This is so whether such expectations take the form of general characteristics used by traditional normalization theories (e.g., formant ranges or vocal tract length) or the more detailed individual apparent-speaker characteristics of exemplar-oriented models.

However, although the indirect effect of f0 on vowel quality seems to be the more salient one, f0 still appears to exert a significant direct effect on vowel quality. The variables we used to measure apparent-speaker characteristics, speaker size and maleness, were, in effect, surrogates for listener-internal latent variables that specify whatever speaker information directly affects vowel quality. It is possible that the apparently direct effect of f0 on vowel quality might actually be due to the fact that our indices of apparent-speaker characteristics (speaker size and maleness) are not sufficient to fully approximate the true values of the relevant internal variables. However, the results we have presented strongly support a theory of vowel perception in which the presumed identity of the speaker plays an important role in the determination of vowel quality. A more elaborate form of latent variable modeling and/or a better set of behavioral instruments relating to relevant judgments of apparent-speaker characteristics might elucidate this question.

In the introduction we suggested the normalization process was being approached as a black box system where we would not seek to define the exact internal working of the process but simply to infer what information plays a significant role in the system's transfer characteristics. At this point it seems fair to say that both f0 and apparent-speaker characteristics play a role in this process in a manner broadly consistent with an indirect model of speaker normalization. However, the precise mechanisms by which these factors operate remains to be determined.

APPENDIX

The negative partial correlation observed (Sec. III A, Table III) between maleness and speaker size judgments is at first glance rather puzzling. However, on further investigation it is clear that there are reasonable explanations for this,

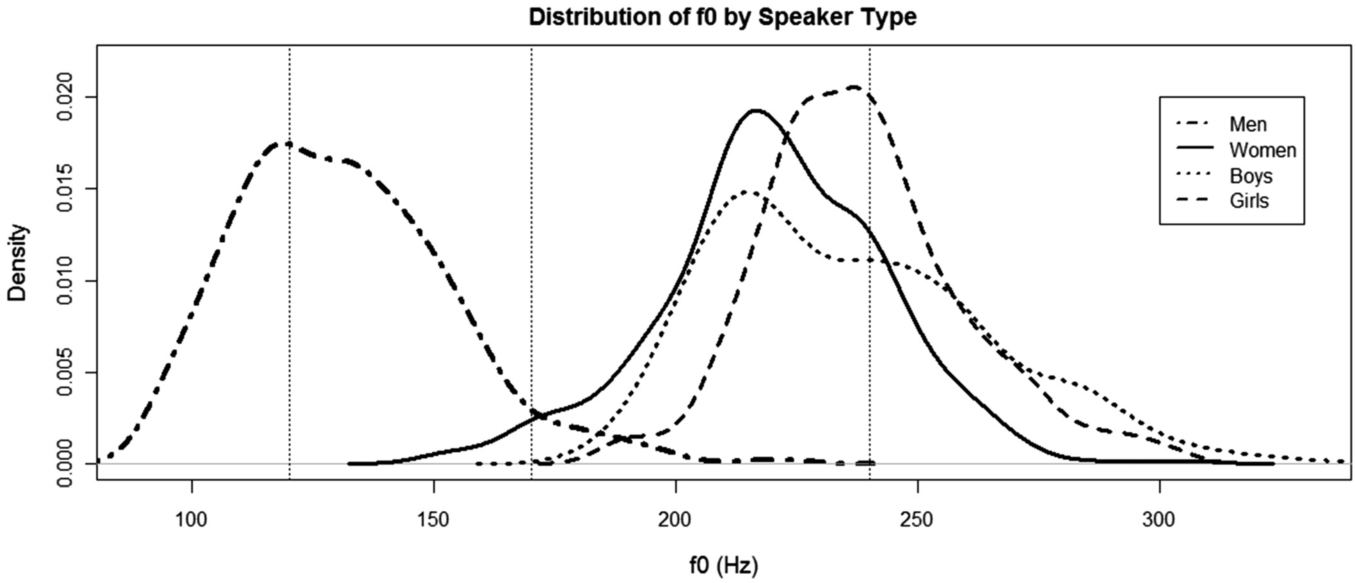

FIG. 4. Kernel density plots for the f0 measurements in the data of Hillenbrand *et al.*, 1995. The vertical lines represent the three f0 levels used in the current experiment.

TABLE V. Percentage of individual vowels (within each speaker group) from the individual data of Hillenbrand *et al.* (1995) that have f0 values exceeding the frequencies used in the current experiment.

	Male adult	Female adult	Male child	Female child
High f0 (240 Hz)	0%	18.6%	40.1%	40.4%
Mid f0 (170 Hz)	5.4%	97.4%	100%	100%
Low f0 (120 Hz)	64.3%	100%	100%	100%

which do not affect the interpretation of the other relationships found.

One possible explanation relates to how the speaker size ratings were used by listeners. There are two ways that immediately spring to mind: First, absolutely across genders; and second, relatively within genders. In the absolute usage, listeners may have used a single scale, roughly proportional to overall speaker body length (or body mass or volume). In this case, the negative correlation between gender and size judgments would be difficult to explain without bringing further evidence to bear. But in the relative, within-gender usage, a negative partial correlation might readily result. For example, suppose a listener decides a stimulus was an /æ/ that sounded as if it was spoken by an individual who was about 165 cm in height, but whose gender was not immediately obvious. If the listener decided ultimately it was a male, they might choose a relatively small size rating because 165 cm is fairly short for a male. However if the listener decided it was a female, they might choose a relatively large size rating, because 165 cm is moderately tall for a female. Suppose on a second replication, the listener made the same assessment of the stimuli, but decided the opposite gender. Cases such as this would contribute to a negative correlation between maleness and speaker size judgments after controlling for all the stimulus factors and vowel judgment.

Another possible explanation involves consideration of the synthetic stimuli in relation to the distribution of acoustic properties measured from natural speech within and across genders. We focus here on f0, which appears to be the strongest determinant of perceived speaker size and maleness (see Sec. III A). The distribution of speaker-size responses with respect to the f0 levels used in this experiment will be discussed in reference to data collected by Hillenbrand *et al.* (1995) (vowel data available from http://homepages.wmich.edu/~hillenbr/). This data set consisted of vowels produced by 50 adult males and females, 29 male children, and 21 female children (all children were between 10–12 years old). Figure 4 presents the distribution of f0s in this data divided by speaker type, while Table V presents the percentage of tokens from each distribution that exceed the f0 levels used for stimuli in this experiment.

Although no adult males in the Hillenbrand data have an f0 as high as 240 Hz, 40.1% of male children's vowels are at least this high. This means that throughout the course of their lives, male speakers have f0s that change from values near those of the high f0 condition to values near those of the low f0 condition. Presumably, at some point during this change they may also have speaking f0s near the mid f0 condition (since this lies between the low and high f0 levels). This naturally leads to a condition in which the f0 levels can be judged as appropriate for a wide range of male speakers, from large to small.

On the other hand, the high f0 level used is close to the average adult female speaking f0 in the Hillenbrand data. As a result, a female speaker with an f0 of 240 Hz may be interpreted as being near normal adult size. The speaking f0 of a typical female speaker does not drop as far as the mid f0 level and would certainly not reach the lowest f0 level. Given that lower f0s are typically associated with larger speakers, vowels with low and mid f0 levels that were interpreted as coming from a female speaker may have led to the impression that the speaker was much larger than the average adult female. The net result of this is that, for any given f0 level, a perceived male speaker will be judged to be smaller than a perceived female speaker (relative to average for that gender).

These facts are reflected in the distribution of speaker size responses when grouped by f0 level and gender response as is shown in Fig. 5. A low f0 level led to the perception of a slightly above average (over all responses) male. Increases in f0 levels lead to movement of the mass of the distribution towards the lower end of the scale, so that speaker size responses shifted from slightly over the middle to the bottom of the scale. However, when listeners reported hearing a female speaker, the shift in size responses was much more

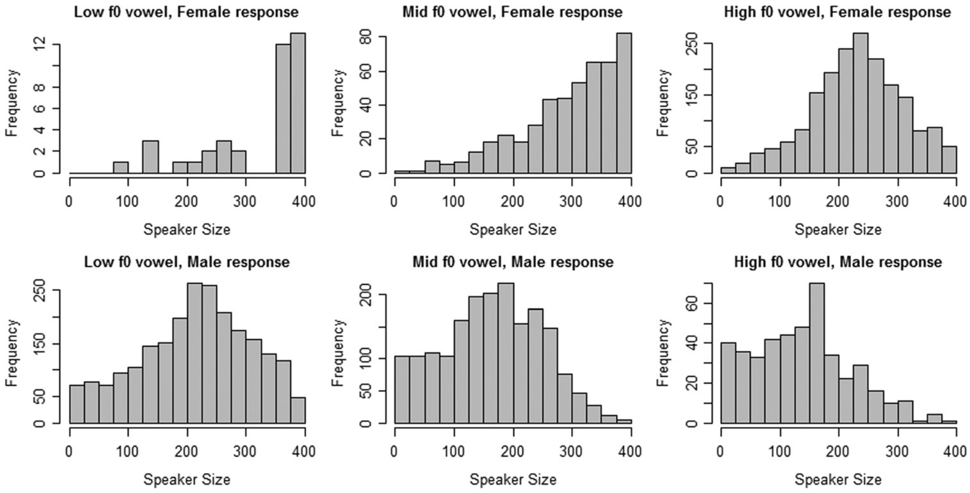

FIG. 5. Distribution of speaker size responses for each combination of gender response and vowel f0. Note that each panel has a different *y*-axis range.

limited. In the rare cases where listeners heard a female speaker with a low f0, the speaker was reported as very large, usually near the very top of the size scale. As f0 levels increase, the size responses for perceived female speaker also move down the scale, but they settle somewhere around the middle rather than towards the lowest extreme.

The relationship exhibited in these graphs is consistent with a negative partial correlation between maleness and speaker size. The within-participant partial correlation was calculated between speaker size and maleness after controlling for f0 only. The average partial correlation was -0.437 ($t=-9.78$, $\text{df}=18$, $p<0.00001$), which is very similar to the -0.475 value reported for the partial correlation between speaker size and maleness, controlling for all other factors (reported in Sec. III A). This indicates that the association between perceived maleness and a (relatively) smaller perceived speaker remains after controlling for the rest of the variables considered in our analysis (F1, F3+, vowel openness).

[1] There is a general correlation between speaker size and average vocal tract length (Fitch and Giedd, 1999; Hollien *et al.*, 1994) across genders and speakers of all ages and sizes. However, after controlling for age and gender, there is no correlation between speaker height and weight and estimated vocal tract length (van Dommelen and Moxness, 1995; Collins, 2000; Gonzalez, 2004). There is also no correlation between speaker height and weight and average speaking f0 when controlling for gender and age (Lass and Brown, 1978; Kunzel, 1989). As a result, it is not surprising that several studies have found that listeners are not very good at judging the actual size of a speaker solely on the basis of their speech (van Dommelen, 1993; Collins, 2000; Rendall, 2007). Although listeners are not very accurate when estimating speaker size, their estimates, both correct and incorrect, have been found to be fairly consistent both within and between listeners (van Dommelen and Moxness, 1995; Lass *et al.*, 1980; Collins, 2000), and are strongly influenced by both f0 and the FFs (Collins, 2000; Smith and Patterson, 2005; Smith *et al.*, 2005; Rendall *et al.*, 2007).

[2] However, the experiment of Smith *et al.* used only five phonetically dissimilar seed vowels /i, e, a, o, u/ from a single speaker. In experiments using similar vocoding techniques, but 12 vowel categories and several speakers, Assmann and Nearey (2008) found considerable variation in vowel identification rates as a function of the relation between spectrum-envelope scaling and f0.

[3] Other sources of variation, such as vocal effort, may also affect the relation between stimulus properties and perceived vowel quality (see Traunmüller, 1994). However, for monosyllabic stimuli in mixed-speaker type presentation with a simple falling intonation pattern, it seems unlikely that these potential sources of variance would have much effect. Furthermore, any effect they did have would simply tend to weaken any relations of the kind we are studying here and should not add any spurious correlations.

[4] An octave increase in Hz, is equal to 0.693 ln (Hz). This times the 0.31 scale factor gives us 0.214, the difference between the FFs of males and females.

[5] Our usage of the term "frame of reference" here and in the discussion refers to the formant space that is likely to be used by a speaker. This usage is consistent with the tradition of Joos (1948), Ladefoged and Broadbent (1957), and Nearey (1989).

[6] All variable pairs except speaker size and vowel are significant at less than a Šidak adjusted one-tailed (in the expected direction) single test level of $p=0.00874$ for a family size of 12.

[7] Apparent-speaker characteristics were also accounted for in the larger model, however, these should not affect the outcome according to f0-free hypotheses.

Assmann, P. F., Dembling, S., and Nearey, T. M. (**2006**). "Effects of frequency shifts on perceived naturalness and gender information in speech," in *Proceedings of the Ninth International Conference on Spoken Language Processing*, Pittsburgh, PA, September 17–21, **2006**, pp. 889–892.

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

Assmann, P. F., and Nearey, T. M. (**2008**). "Identification of frequency-shifted vowels," J. Acoust. Soc. Am. **124**, 3203–3212.

Bachorowski, J., 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.

Boersma, Paul, and Weenink, David (**2009**). "Praat: Doing phonetics by computer," Version 5.1.01, http://www.praat.org/ (date last viewed October 2008).

Collins, S. A. (**2000**). "Men's voices and women's choices," Anim. Behav. **60**, 773–780.

Fant, Gunnar (**1960**). *Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations* (Mouton, The Hague), pp. 107–138.

Fitch, W. T., and Giedd, J. (**1999**). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," J. Acoust. Soc. Am. **106**, 1511–1522

Fujisaki, H., and Kawashima, T. (**1968**). "The roles of pitch and higher formants in the perception of vowels," IEEE Trans. Audio Electroacoust. **AU-16**, 73–77.

Glidden, C., and Assmann, P. F. (**2004**). "Effects of visual gender and frequency shifts on vowel category judgments," Acoust. Res. Lett. Online **5**, 132–138.

Gonzalez, J., (**2004**). "Formant frequencies and body size of speaker: A weak relationship in adult humans," J. Phonetics **32**, 277–287.

Gottfried, T. L., and Chew, S. L. (**1986**). "Intelligibility of vowels sung by a countertenor," J. Acoust. Soc. Am. **79**, 124–130.

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

Hillenbrand, J. M., and Houde, R. A. (**2003**). "A narrow band pattern-matching model of vowel perception," J. Acoust. Soc. Am. **113**, 1044–1055

Hollien, H., Green, R., and Massey, K. (**1994**). "Longitudinal research on adolescent voice change in males," J. Acoust. Soc. Am. **96**, 2646–2653.

Irino, T., and Patterson R. D. (**2002**). "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform," Speech Commun. **36**, 181–203.

Johnson, Keith. (**1990**). "The role of perceived speaker identity in f0 normalization of vowels," J. Acoust. Soc. Am. **88**, 642–654.

Johnson, Keith, Strand, Elizabeth A., and Mariapaola D'Imperio. (**1999**). "Auditory-visual integration of talker gender in vowel perception," J. Phonetics **27**, 359–384.

Johnson, Keith. (**2005**). "Speaker normalization in speech perception," in *The Handbook of Speech Perception*, edited by D.B. Pisoni and R. Remez (Blackwell, Oxford), pp. 363–389.

Joos, M. (**1948**). "Acoustic phonetics," Language **24**, 1–136.

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

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

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**). "Correlational study of speakers' heights, weights, body surface areas, and speaking fundamental frequencies," J. Acoust. Soc. Am. **63**, 1218–1220

Lass, N. J., Phillips, J. K., and Bruchey, C. A. (**1980**). "The effect of filtered speech on speaker height and weight identification," J. Phonetics **8**, 91–100.

Lehiste, I., and Meltzer, D. (**1973**). "Vowel and speaker identification in natural and synthetic speech," Language Speech **16**, 356–364.

Lorch, R. F., and Myers, J. L. (**1990**). "Regression analyses of repeated measures data in cognitive research," J. Exp. Psychol. Learn. Mem. Cogn. **16**, 149–157.

Mendoza, E., Valencia, N., Muñoz, J. and Trujillo, H. (**1996**). "Differences in voice quality between men and women: Use of the long-term average spectrum (LTAS)," J. Voice **10**, 59–66.

Miller, J. D. (**1989**). "Auditory-perceptual interpretation of the vowel," J. Acoust. Soc. Am. **85**, 2114–2134.

Miller, R. L. (**1953**). "Auditory tests with synthetic vowels," J. Acoust. Soc. Am. **25**, 114–121.

Nearey, T. M. (**1989**). "Static, dynamic, and relational properties in vowel perception," J. Acoust. Soc. Am. **85**, 2088–2113.

Nearey, T. (**1992**). "Context effects in a double-weak theory of speech perception," Lang. and Speech **35**, 153–172.

Nearey T. M., and Assmann P. F. (**2007**). "Probabilistic 'sliding template' models for indirect vowel normalization," in *Experimental Approaches to Phonology*, edited by M. J. Solé, P. S. Beddor, and M. Ohala (Oxford University Press, Oxford), pp. 246–269.

Perry, T. L., Ohde, R., N., and Ashmead, D. N. (**2001**). "The acoustic bases for gender identification from children's voices," J. Acoust. Soc. Am. **109**, 2988–2998.

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

Price, P. J. (**1989**). "Male and female voice source characteristics: Inverse filtering results," Speech Commun. **8**, 261–277.

Rendall, D., Vokey, J. R., and Nemeth, C. (**2007**). "Lifting the curtain on the Wizard of Oz: Biased voice-based impressions of speaker size," J. Exp. Psychol. Hum. Percept. Perform. **33**, 1208–1219.

Slawson, A. W. (**1968**). "Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency," J. Acoust. Soc. Am. **43**, 87–101

Smith, David R. R., and Roy, D. Patterson (**2005**). "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age," J. Acoust. Soc. Am. **118**, 3177–3186

Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H., and Irino, Toshio (**2005**). "The processing and perception of size information in speech sounds," J. Acoust. Soc. Am. **117**, 305–318.

Strand, Elizabeth A. (**2000**). "Gender stereotype effects in speech processing," Ph.D. dissertation, The Ohio State University.

Syrdal, A. K., and Gopal, H. S. (**1986**). "A perceptual model of vowel recognition based on the auditory representation of American English vowels," J. Acoust. Soc. Am. **79**, 1086–1100.

Traunmüller, H. (**1994**). "Conventional, biological, and environmental factors in speech communication: A modulation theory," Phonetica **51**, 170–183.

Van Borsel, J., Janssens, J., and De Bodt, M. (**2009**). "Breathiness as a feminine voice characteristic: A perceptual approach," J. Voice **23**, 291–294.

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

van Dommelen, W. A. (**1993**). "Speaker height and weight identification: A re-evaluation of some old data," J. Phonetics **21**, 337–341.