Training listeners to report the acoustic correlate of formant-frequency scaling using synthetic voices

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

(Received 23 May 2012; revised 30 November 2012; accepted 12 December 2012)

The vocal tract length of a speaker is the primary determinant of the range of formant frequencies (FFs) produced by that speaker. Listeners have demonstrated sensitivity to the average FFs produced by voices, for example, in estimating the relative heights of two speakers based on their speech. However, it is not known whether they can learn to identify voices based on the acoustic characteristic associated with the average FFs produced by a voice (this characteristic will be referred to as FF-scaling). To investigate this, a series of vowels corresponding to voices that differed in their average *f*0 and/or FF-scaling were synthesized. Listeners ($n = 71$) were trained to identify these voices using a training procedure where, for each trial, they heard the vowels representing a voice and then had to identify the stimulus voice from among a series of candidate voices that differed in terms of their FF-scaling and/or their *f*0. Results indicate that listeners can identify voices on the basis of FF-scaling quite accurately and consistently after only a short training session and that, although *f*0 weakly influences these estimates, they are most strongly determined by the stimulus FFs.
[http://dx.doi.org/10.1121/1.4773858]

PACS number(s): 43.71.Bp, 43.71.An, 43.71.Es [JMH] Pages: 1065–1077

I. INTRODUCTION

Since the first acoustic studies in the 1950s, variation in the acoustic properties of vowels of different speakers has typically been discussed in terms of their fundamental frequency (*f*0) and formant frequencies (FFs). The scaling of *f*0 and FF ranges has also figured prominently in parametric synthesis of voices simulating speakers of different sizes, genders, and age groups (Klatt and Klatt, 1990). Although the perception of *f*0 has been extensively studied, the perception of the acoustic characteristic associated with the range of formant frequencies produced by different speakers is not as well understood. In the sections that follow, a case will be made for the importance of this acoustic characteristic that we will call formant-frequency scaling (or FF-scaling), in the listener's assessment of apparent speaker characteristics (i.e., the indexical characteristics of the speaker inferred by the listener), and the perception of vowel quality. Furthermore, we suggest that the importance of FF-scaling in both vowel perception and the determination of apparent speaker characteristics may explain the relationship between these processes observed in several previous experiments.

In the discussion below, we will be adopting the uniform scaling hypothesis as a working assumption. Uniform scaling proposes that a set of phonetically equivalent vowels produced by two speakers of the same dialect are (on average) relatable to each other by a single multiplicative parameter. Although there is some controversy about this in the literature (see the Appendix), in practice it leads to reasonably good approximations of systematic speaker variability (Nearey, 1978; Nearey and Assmann, 2007; Turner *et al.*, 2009). The scaling parameter (i.e., FF-scaling) is related to speaker vocal-tract length and determines the relative scaling applied to the formant-pattern of a given vowel by the vocal tract of the speaker.

A. FF-scaling and apparent speaker characteristics

Because of their dependence on the anatomy of the speaker, the average *f*0 and FFs produced by a speaker covary with some prominent speaker characteristics. Men tend to have lower *f*0s than women, and children tend to have higher *f*0s than adults of the same gender so that *f*0 correlates strongly to speaker height across all speakers (Hollien *et al.*, 1994). The average FFs produced by a speaker will be most strongly determined by that speaker's vocal-tract length, with longer vocal tracts producing lower FFs overall, and shorter vocal tracts producing higher FFs overall (Fant, 1960). There is a strong positive correlation between speaker height and speaker vocal-tract length (Fitch and Giedd, 1999) so that, in general, larger speakers have lower FF-scalings overall than smaller speakers (Lee *et al.*, 1999; Peterson and Barney, 1952). Consequently, the *f*0 and FFs of a vowel represent two potentially different streams of information arising from two acoustically distinct origins, each of which may be used by listeners to estimate speaker characteristics, such as height or gender.

Speakers may be divided into four general speaker classes based on two dichotomies: child vs adult and male vs female. If speakers are sorted to fit into one of these categories, then the average *f*0 and FF-scaling differences between speaker classes can be quite large. For example, an automatic classifier can predict the gender of an adult speaker with up to 98% accuracy using only information regarding

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

the FF-scaling and *f*0 that characterize that voice (Hillenbrand and Clark, 2009). However, the correlation between speaker height and voice characteristics (FF-scaling and *f*0) within a single class (e.g., adult males) is unreliable, particularly for adult speakers who have reached a stable height. There is no significant correlation between adult speaker height and average *f*0 after controlling for gender (Hollien *et al*., 1994; Gonzalez, 2004; Lass and Brown, 1978; Collins, 2000; van Dommelen and Moxness, 1995). It has similarly been reported that there is no significant correlation between adult speaker height and FF-scaling after controlling for gender (Collins, 2000; van Dommelen and Moxness, 1995), or that the correlation is weak[1] (Gonzalez, 2004).

Given that the relationship between the acoustic properties of the vowels produced by a speaker and that speaker's height is weak within a speaker class, it is not surprising that listeners are not able to accurately estimate speaker height based on a speaker's *f*0 and FF-scaling when speaker class is controlled, for example, by presenting listeners with speech from adult speakers only (van Dommelen and Moxness, 1995; Collins, 2000; Rendell *et al*., 2007). Despite the inability of listeners to arrive at *veridical* estimates of speaker size based on speech samples, listeners typically arrive at *consistent* judgments regarding a speaker's size, both within and across listeners (van Dommelen and Moxness, 1995; Collins, 2000; Smith and Patterson, 2005; Rendell *et al*., 2007).

The manner in which listeners estimate speaker height has been investigated by presenting listeners with speech sounds that vary in terms of *f*0 and FF-scaling, but with a fixed phonetic content, and asking listeners to assess the absolute or relative heights of speakers. This has been done using synthetic vowels (Fitch, 1994) and modified natural-speech (Ives *et al*., 2005; Smith and Patterson, 2005; Smith *et al*., 2005; Rendell *et al*., 2007). Results indicate that these judgments are informed by jointly considering the FF-scaling and *f*0 of a voice (Fitch, 1994; van Dommelen and Moxness, 1995; Smith and Patterson, 2005), where progressively lower FF-scalings and/or progressively lower *f*0s suggest a progressively larger speaker.

Most listeners are familiar with the concept of pitch, and it is known that they can make overt judgments of pitch that relate to the relative *f*0 level of different voices (Honorof and Whalen, 2005). It is not clear, however, whether there exists any separable perceptual dimension that corresponds closely to FF-scaling that listeners might learn to report. Since this putative perceptual dimension[2] has no name that we know of, we will refer to it tentatively as the perceptual FF-scale estimate, or pFF-scaling, to keep it distinct from the acoustic FF-scaling used to create the stimuli used in the experiment to be outlined below.

To date, experiments involving listener responses to variations in the FF-scaling of voices have focused on the estimation of speaker characteristics (e.g., gender, body size), which are determined by jointly considering voice *f*0 and FF-scaling. For example, a common methodology (Fitch, 1994; Smith and Patterson, 2005) involves creating a set of stimuli with fixed phonetic content, which span an *f*0 × FF-scaling space (as in Fig. 2). Listeners are then presented with these stimuli in a random order and, for each trial, are asked to estimate some speaker characteristic, for example, the speaker's height or gender. By comparing the rated heights of voices at different points within an *f*0 by FF-scaling space, researchers may investigate the relative contribution of each cue to such judgments via linear regression. Although this methodology can shed light on the manner in which speaker characteristics are determined by jointly considering voice *f*0 and FF-scaling, they cannot provide information about listeners' use of any perceptual dimension or mechanism that specifically follows physical variation in FF-scaling as such.

For example, consider two voices with the same *f*0 and source characteristics, one of which has a lower FF-scaling than the other. If one listener reports hearing a male for the low FF-scaling voice, and a female for the high FF-scaling voice, it is reasonable to infer that they are responding to a change in voice FF-scaling. However, if a second listener reports that both voices appear to represent male speakers, this does not entail that the listener fails to notice the difference in FF-scaling. Rather, the second listener may have a higher threshold for a change in apparent speaker gender, or they may attribute the change in FF-scaling to a change in size-within-gender or any number of factors (including, for some formant patterns at least, differences in vowel quality whether categorical or graded). In short, the collection of judgments of apparent speaker characteristics does not allow researchers to directly investigate the perception of FF-scaling or its putative perceptual counterpart pFF-scaling. As discussed in Sec. I C, if listeners are able to provide perceptual judgments that correlate well with FF-scaling, such judgments could be a valuable source of information in the evaluation of perceptual theories related to vowel-normalization.

B. FF-scaling, normalization, and vowel perception

Several theories of human vowel perception involve the estimation of a speaker-dependent formant-space as a frame of reference used to interpret the vowels produced by a speaker (Joos, 1948; Ladefoged and Broadbent, 1957; Ainsworth, 1975; Nearey, 1978: Nearey, 1989: Nearey and Assmann, 2007). The speaker-dependent formant-space need only be detailed enough so that a listener knows roughly what FFs to expect for a given vowel category when produced by that speaker. The listener then identifies vowels by considering the FFs of a vowel sound relative to expected FFs for each vowel category, rather than by considering the FFs in an absolute manner. This general hypothesis is typically referred to as *speaker normalization*. To the extent that variation in formant-spaces across speakers can be accounted for by a single parameter (i.e., FF-scaling), the process of speaker normalization can be thought of as centering around the estimation of an appropriate FF-scaling with which to identify vowels produced by that speaker.

This insight underlies the log-mean normalization method proposed in Nearey (1978). It has been used routinely for decades in sociophonetic studies by Labov and his colleagues, where it has been found to be effective for preserving relatively subtle systematic differences between

dialects and sociolects while largely removing effects of vocal tract length (Labov *et al.*, 2006, p. v). This method calculates the log-mean FF produced by a speaker across their entire vowel system, a measure which should be strongly correlated with speaker FF-scaling, and subtracts this value from the log-transformed formant frequencies produced by a speaker. In effect, this method centers the vowel spaces of different speakers along the primary axis of variation between speakers (i.e., $\ln F1 = \ln F2$; see the Appendix) and, consequently, allows variation in FFs to be interpreted more directly as evidence of differences in vowel quality (as opposed to simply being a result of differences in speaker vocal-tract length).

Consider Fig. 1, which presents the Peterson and Barney (1952) vowel data (also presented in Fig. 3 in the Appendix). In this figure, FFs have been normalized using the log-mean method of Nearey (1978). As seen in Fig. 1, this process greatly reduces the between-category overlap between vowel categories relative to the raw FFs (presented in Fig. 3). Furthermore, the major axes of the ellipses representing the different vowel categories are no longer primarily aligned with the $\ln F1 = \ln F2$ axis as they are for unnormalized data (see the Appendix). In fact, whereas variation along this axis accounted for 80.6% of the variance in FFs in the unnormalized FFs (a ratio of nearly 4/1), after normalization variation along this axis accounts for only 52.9% of variation, on average indicating an essentially equal distribution in variation along $\ln F1 = \ln F2$ and the orthogonal axis.

Nearey and Assmann (2007) present an empirical example of the potential usefulness of an FF-scaling estimate in vowel perception. They describe an automatic vowel classifier that identifies vowels based on the FFs of a vowel, a dialect-specific template indicating the positions of vowel categories in a formant-space, and a speaker-specific FF-scaling estimate. The speaker-specific FF-scaling estimate, which they refer to as ψ, is the log-mean of the FFs produced by a speaker across their entire vowel system (Method 1). Although this may not be a realistic model of human vowel perception (since human listeners can identify vowels without hearing a speaker's entire vowel system), it offers a limiting-case for the usefulness of a speaker-dependent FF-scaling estimate in vowel perception. This classifier correctly identifies vowels from the Hillenbrand *et al.* (1995) data set in nearly 93% of cases, and vowels from the Assmann and Katz (2000) data set in 82% of cases. For both data sets, performance compares favorably to that of human listeners, reported at 95.4% (Hillenbrand *et al.*, 1995) and 84% (Katz and Assmann, 2001), respectively.[3]

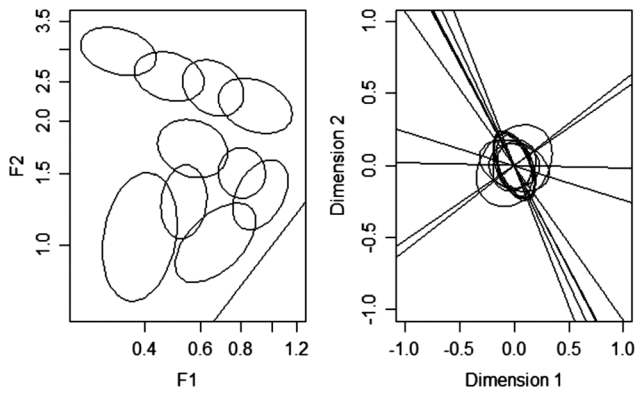

FIG. 1. In the left panel, ellipses enclosing two standard deviations of the Peterson and Barney (1952) vowels are presented. Vowels have been normalized-using the log-mean normalization method of Nearey (1978). $F1$ and $F2$ are presented as the ratio of each formant frequency to the geometric mean $F1$-$F2$-$F3$ frequency produced by each speaker across their whole vowel system. The line is a line parallel to $\ln F1 = \ln F2$. In the right panel, all formant frequencies have been log-transformed and centered within-category so that vowel-category means are at the origin. All points have been rotated 45 deg clockwise so that the $\ln F1 = \ln F2$ line is now parallel to the x-axis (Dimension 1), while Dimension 2 represents the orthogonal axis. The lines indicate the major axes of the vowel-category ellipses, which no longer vary primarily along the $\ln F1 = \ln F2$ axis (Dimension 1).

Although a speaker-dependent FF-scaling estimate may play an important role in vowel perception, the listener does not have direct access to the speaker's true FF-scaling, and must estimate this value. Both Nearey and Assmann (2007) and Turner *et al.* (2009) have emphasized that since the uniform scaling hypothesis entails that productions between speakers of the same vowel differ by a single multiplicative parameter (i.e., FF-scaling), identifying a vowel sound will yield an estimate of the speaker-specific parameter (i.e., pFF-scaling), since listeners may infer the speaker's FF-scaling given the observed formant frequencies. This is analogous to the manner in which identifying a visual object of a known physical size yields an estimate of its distance from the observer. In this view of vowel perception, the speaker-dependent FF-scaling estimate, pFF-scaling, might be thought of as a derived perceptual property, which a listener constructs in establishing a speaker-dependent formant-space with which to interpret a speaker's vowels.

C. FF-scaling, vowel perception, and apparent speaker characteristics

Because of the potential importance of FF-scaling estimates in human vowel normalization, the ability to collect them from listeners may help clarify unresolved issues in the study of speech perception. For example, previous studies have found that vowel quality shifts can be induced by manipulating vowel $f0$, or the $f0$ of a preceding carrier phrase (Miller, 1953; Fujisaki and Kawashima, 1968; Slawson, 1968; Nearey, 1989; Johnson, 1990). Similar effects have been observed by pairing vowel sounds with male or female faces (Glidden and Assmann, 2004), or simply by telling listeners that the speaker is of a certain gender (Johnson *et al.*, 1999). Johnson (1990, 2005) and Johnson *et al.* (1999) have suggested that $f0$ affects vowel quality primarily indirectly, by affecting apparent speaker characteristics, rather than by being directly involved in the specification of vowel quality.

In terms of a general theory of speaker normalization, $f0$ is expected to affect perceived vowel quality primarily by informing the speaker-dependent formant-space used by the listener to interpret the vowels of a speaker. Apparent speaker gender is expected to affect perceived vowel quality in a similar manner. For example, if a vowel is presented with a high pitch, a listener may assume that the speaker is a female and may assume a formant-space appropriate for a female speaker. If a vowel with the same FFs were presented

with a low pitch, the listener may assume a male speaker, and a formant-space appropriate for a male, which may lead to differences in perceived vowel quality. This may be contrasted with the direct effect of a change in $F1$, for example, which would be expected to result in a change in vowel quality even within-speaker.

Barreda and Nearey (2012a) report the results of an experiment that offers strong support for Johnson's hypothesis. Listeners were presented with a series of vowels that differed in their FFs and $f0$ and, for each trial, were asked to report vowel quality and two apparent speaker characteristics. The speaker characteristics they were asked to report were speaker gender (male or female) and speaker size (using a continuous scale that they were instructed to use as they saw fit). Results indicate that although $f0$ can exert a strong influence on perceived vowel quality, this effect is greatly diminished (but still significant) if apparent speaker characteristics are accounted for. This was taken as an indication that although $f0$ is strongly related to perceived vowel quality, its effect is mostly achieved by suggesting apparent speaker characteristics to the listener. Furthermore, apparent speaker gender had a significant effect on perceived vowel quality, and apparent speaker size (controlling for gender) had a marginally significant effect[4] on vowel quality, even after controlling for the acoustic characteristics of the vowel sound.

Although experiments such as Johnson (1990), Johnson *et al*. (1999), Glidden and Assmann (2004), and Barreda and Nearey (2012a) used speaker characteristics such as speaker gender to investigate the process of speaker normalization, none of these authors suggest that speaker gender is directly involved in the specification of vowel quality in the same way that the formants are. Rather, these experiments might be interpreted as using apparent speaker characteristics as surface variables to investigate the latent variable of interest, the FF-scaling estimate for a voice on the part of the listener. Because of the strong and consistent association listeners make between FF-scaling and perceived speaker size and gender (outlined in Sec. I A), experimenters might reasonably infer that if listeners indicate that a speaker is an adult male, they will also expect a relatively lower FF-scaling than if the speaker were an adult female. Thus, controlling for apparent speaker characteristics, as in Barreda and Nearey (2012a), can be viewed as indirectly attempting to control for a latent estimated FF-scaling, while affecting apparent speaker gender as in Glidden and Assmann (2004) might be viewed as an attempt to influence implicit, listener-internal FF-scaling estimates.

A more direct approach to experiments investigating the direct and indirect effects of acoustic cues on vowel quality would be to collect overt FF-scaling judgments from listeners in experiments designed to investigate specific questions. If this could be done, researchers would not need to rely solely on speaker characteristics that, although they may strongly co-vary with speaker FF-scaling, may do so only in a complex, derivative way. Furthermore, specific hypotheses about the possible role of FF-scaling estimates in vowel perception could be tested in a more direct manner.

D. Rationale for the current experiment

In Secs. I A–I C, we have established that the formant patterns produced by speakers of different sizes vary primarily in terms of a single, multiplicative parameter, which we refer to as FF-scaling. Because of its strong relationship to speaker vocal-tract length, this acoustic characteristic is closely related to salient apparent speaker characteristics such as size and gender. Listeners may take advantage of this co-variation, and use FF-scaling information to infer apparent speaker characteristics from the speech signal. We have outlined a case for the potential centrality of information related to speaker FF-scaling in human vowel perception in terms of a general process of speaker normalization. Finally, we have suggested that the effect of some apparent speaker characteristics on perceived vowel quality may occur by means of influencing the listener's speaker-dependent FF-scaling estimate.

Although the line of reasoning summarized in the previous paragraph has extensive experimental and theoretical support, the perception of speaker FF-scaling is not well understood. Given that our position on the process of vowel perception centers around a speaker-dependent FF-scaling estimate, it is incumbent on us to demonstrate that listeners are able to identify voices that differ according to this acoustic characteristic, and to investigate the nature of a possible pFF-scaling perceptual dimension.

Despite the potential usefulness of obtaining voice FF-scaling estimates from listeners, no previous experiment has focused on training listeners to directly report this property. The purpose of this experiment is to investigate the extent to which listeners can learn to distinguish and identify voices that vary in both average $f0$ and FF-scaling. The experiment to be outlined here adopts a similar stimulus design to that employed in Fitch (1994), and Smith and Patterson (2005), where listeners are presented with a series of stimuli that span an $f0 \times$ FF-scaling space but have a fixed phonetic content. However, instead of a rating-scale judgment of a specific speaker characteristic, listeners are trained to provide absolute identifications of each voice presented from a discrete set of alternatives in a two-dimensional display corresponding to an $f0 \times$ FF-scaling space. In doing so, listeners will provide what can be viewed as estimates of voice $f0$ and voice FF-scaling independently for each dimension,[5] rather than providing a measure (such as judged size or gender) that is likely to involve joint consideration of the two properties.

This experiment also seeks to investigate the feasibility of collecting FF-scaling judgments from listeners in varying experimental conditions. Future experiments investigating the manner in which listeners estimate voice FF-scaling may require listeners to report voice $f0$, or they may require listeners to disregard it, depending on the specific question being addressed. To investigate whether disregarding stimulus $f0$ results in a significant change in the consistency with which listeners report voice FF-scaling, the ability of listeners to report voice FF-scaling will be tested in two conditions. In the first of these, listeners will be asked to report FF-scaling and $f0$ for each trial. In the second condition,

listeners will be asked to report FF-scaling only, and disregard stimulus *f*0.

There are three general possible outcomes, each of which has different implications for the manner in which human listeners respond to and isolate the FF-scaling of a voice, and for the nature of an acoustic quality such as pFF-scaling. The first possible outcome is that listeners are not able to do this and perform no better than chance in either of the testing conditions. This outcome would be problematic given that listeners have been found to respond to FF-scaling changes in determining apparent speaker characteristics. This outcome might suggest that listeners' representations of voice characteristics are not organized along dimensions related to FF-scaling, that the training paradigm was fundamentally flawed in some way, or, finally, that the task was too difficult given the relatively short training sequence.

The second possible outcome is that listeners are able to report their judgments of FF-scaling with good consistency and accuracy (that is, the judgments are strongly correlated with the physical FF-scaling of the stimuli), and that these judgments are made independently of stimulus *f*0. This outcome would be predicted based on the work by Irino and Patterson (2002), Smith *et al.* (2005), and Turner *et al.* (2006), which have all suggested that the peripheral auditory system processes sounds at an early level, and that this processing segregates information regarding the size of the vocal tract from information regarding the particular configuration of the vocal tract during articulation. The output of this process is expected to be directly available to the listener (which would suggest relatively high performance), and FF-scaling identification should not be influenced by *f*0.

The third possible outcome is that listeners are able to report FF-scaling with a good level of accuracy and consistency, but that these judgments are influenced by stimulus *f*0. This outcome would be predicted by processes similar to Method 6 of the sliding template model (Nearey and Assmann, 2007), which estimates speaker FF-scaling on the basis of the joint distribution of *f*0 and FF-scaling between speakers, and the relative fit of the observed FFs to those expected for each vowel category. Importantly, only a main effect of *f*0 on reported FF-scaling is predicted, where a higher *f*0 should result in a higher reported FF-scaling. This predicted outcome will be shared by any proposed normalization method which seeks to exploit the covariance between FF-scaling and *f*0 between speakers to estimate speaker FF-scaling based on *f*0 (although specific models may predict more complicated patterns of relationships between *f*0 and reported FF-scaling).

II. METHODOLOGY

A. Participants

Listeners were 71 students from the University of Alberta drawn from a participant pool in which undergraduate students take part in experiments in exchange for partial course credit. All participants were students taking an introductory level, undergraduate linguistics course. Before beginning the experiment, all participants filled out a questionnaire in which they indicated their age, gender, native language, any other languages they spoke, and the amount of formal musical training they had received (measured in years). This background information was collected because we thought that prior musical or language experience might influence listeners' ability to perform the experimental tasks successfully. Our reasoning is discussed further in Sec. III.

B. Stimuli

The stimuli consisted of vowel pairs with formant patterns appropriate for the sequence [i æ] (in that order, separated by a pause) spoken by a single speaker. These were constructed to simulate the voices of 15 different synthetic speakers. The vowels associated with these voices varied on the basis of three factors: *f*0 step, FF-scaling step, and the difference in FF-scaling between adjacent FF-scaling steps (this difference will be referred to as ΔFF-scale). FF-scaling level and *f*0 level were within-subjects factors, so that each listener was presented with voices at each combination of *f*0 and FF-scaling steps (3 *f*0 steps × 5 FF-scaling steps). However, ΔFF-scale was a between-subjects factor, so that each listener was only ever presented with voices at a single ΔFF-scale level.

The FFs of vowels representing an FF-scaling step were determined by increasing all of the FFs of the previous step by a fixed percentage (i.e., by a single multiplicative scale factor). The size of the percentage increase between adjacent FF-scaling steps was determined by the ΔFF-scale level. Four different FF-scaling increments were used (7%, 8%, 9%, 10%), resulting in four groups of listeners. For example, for the stimuli for the 9% ΔFF-scale level, the FFs of the vowels of the second FF-scaling step were determined by increasing all of the FFs of the first FF-scaling step by 9%. The FFs of the vowels for the third FF-scaling step were then increased by a further 9% relative to those of the second step (18.81% relative to the first FF-scaling step), and so on.

It is worth noting that the FF-scaling differences used in the construction of the stimuli for this experiment (7%, 8%, 9%, 10%) are close to the estimated just noticeable difference for FF-scaling, estimated to be 7%–8% by Smith *et al.* (2005) and 4%–6% by Ives *et al.* (2005). In both cases, just noticeable differences were estimated using a two-alternative, forced-choice methodology.

Each vowel of the [i æ] stimulus pair was 200 ms in length, and these were separated by 125 ms of silence. Table I presents the initial values for each of the three *f*0 steps. For every stimulus, *f*0 decreased linearly by 10% from the beginning to the end of the vowel. *f*0 levels were the same for all ΔFF-scale levels. Table I also provides the FFs

TABLE I. Initial *f*0 levels for all conditions. Formant frequencies provided are those used for the lowest FF-scaling step vowels in all conditions, corresponding to formant frequencies appropriate for a typical adult male.

	Low	Medium	High
*f*0	110	177	270
	F1	F2	F3
i	280	2148	2755
æ	717	1497	2318

used for the first (lowest-frequency) FF-scaling step for all ΔFF-scale levels. These values were set based on average productions of the same vowels produced by adult male native speakers of the regional dialect. For both vowels, *F*4 was set at 3375 Hz and each formant above *F*4 was 1000 Hz higher than the last, up to the tenth formant. Vowels were synthesized with a variable sampling rate so that the Nyquist frequency fell halfway between the tenth formant and the expected frequency of the eleventh formant given the spacing between formants. The inclusion of higher formants, and the variable sampling rate, were undertaken to avoid inappropriate spectral levels that can readily result when there is uneven distribution of formants near the Nyquist frequency (see Nearey, 1989, Appendix B, for a discussion of some of the issues involved). All vowels were then re-sampled at 22 050 Hz. Figure 2 compares the location of the synthetic voices used in this experiment, for each ΔFF-scale level, to a range of real voices plotted on an $f0 \times$ FF-scaling space.

C. Procedure

A training game reminiscent of the "concentration" or "memory" card game was created to train participants to report FF-scaling independently of *f*0. This game was played on a computer using a specially designed graphical user interface. The game board contained 15 boxes arranged in three rows of five. Each of these boxes was associated with a single voice throughout each participant's experimental session. Voices in the same row had the same *f*0 while voices in the same column had the same FF-scaling. Voice *f*0 increased from top to bottom across rows while voice FF-scaling increased from left to right across columns (in fact, the stimulus voices were arranged on the board in the same manner that they are arranged in Fig. 2). Before beginning the game, participants completed an introductory task in which they were familiarized with all voices. Participants were told that the pitch of voices would increase from bottom to top and that voices differed from left to right in terms of "voice size," which they were told was closely related to speaker size.

The general procedure during the training game was that participants were presented with vowels produced by one of the voices on the board and were asked to indicate the position of the voice within the board by clicking on the box that was associated with it. By locating the voice on the board, participants were, in effect, reporting the FF-scaling and *f*0 levels for the stimulus voice. The game consisted of a series of 11 levels of increasing difficulty. Difficulty was increased between levels by increasing the number of candidate voices available to listeners during each trial. For example, initially listeners were asked to identify a voice from one of two candidates, while in later levels listeners were asked to identify a voice from among all voices in a row, or all voices in two rows. Buttons associated with voices that were candidates for selection in the session were colored blue. Buttons that were not to be considered for selection were the same gray color as the background of the board.

The procedure in each level was as follows: For a trial, listeners were played the vowels [i æ], produced by a single voice. These vowels were always presented in the same order and were separated by 125 ms of silence. Listeners were allowed to replay the vowels as many times as they liked by clicking on a button marked "replay." Listeners then had to indicate the location of the voice on the board by

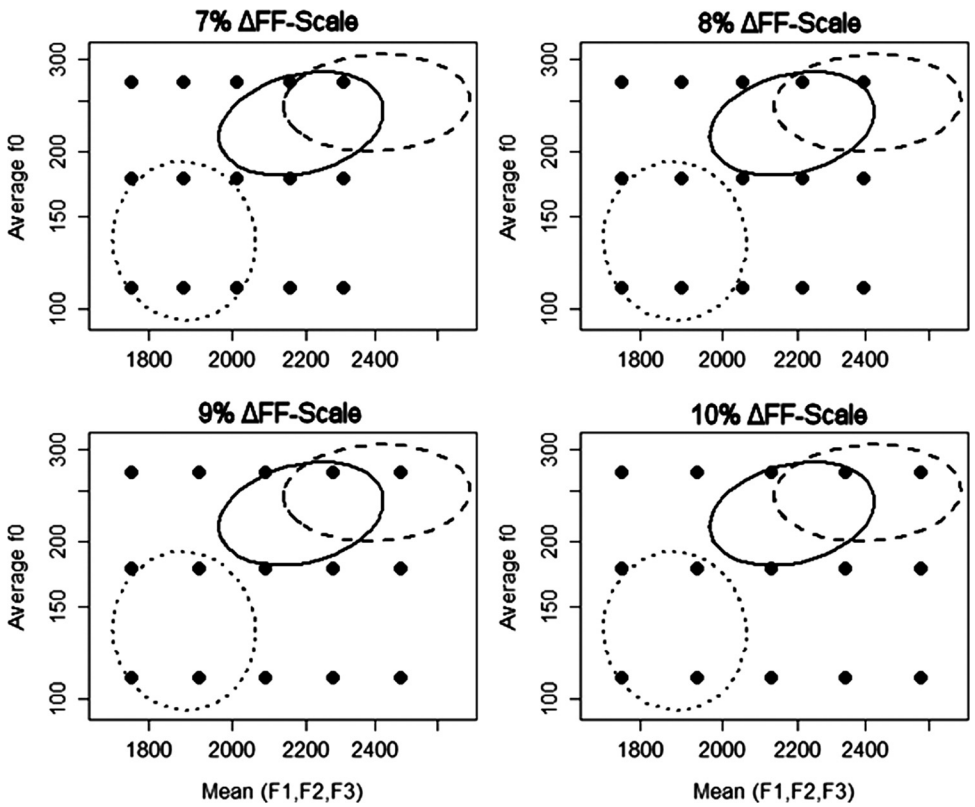

FIG. 2. The *x*-axis indicates the mean of the first three formant frequencies for productions of /i/. Ellipses enclose two standard deviations of the distribution of real voices from data collected by Hillenbrand *et al*. (1995). Ellipses indicate the distribution of voices of adult males (dotted line), adult females (solid line), and children (broken line). The locations of stimulus voices at each ΔFF-scale level are indicated by the filled points.

clicking on one of the blue buttons. When listeners answered correctly, the next pair of vowels played after a 1 s pause, and the process continued until all candidate voices were identified three times each. Voices were presented in a randomized condition, blocked by repetition.

When participants answered incorrectly, the game entered into a special game mode designed to provide the user with feedback, and an opportunity to improve their performance by listening to the voices on the board. In this mode, the correct location of the voice the listener had just heard was indicated by a green box. The box that had been incorrectly selected by the listener was indicated by a red box. Listeners were allowed to listen to all available voices as many times as they liked by clicking on the boxes associated with different voices. When a listener was finished using error mode, they clicked on a button marked "resume," after which the next voice in the round was presented after a 1 s pause.

Longer-term feedback was provided to listeners via a message across the top of the game board, which informed listeners of the percent of trials they had identified correctly within a given level and of the percentage of trials in which they had been within one step at most, in both *f*0 and FF-scaling level, of the correct box. When a level was completed, listeners moved on to the next level in the game by clicking on the button marked "resume." The next level would not begin until the listener clicked on this button. All listeners took part in experimental sessions of a maximum of 1 h in length.

After completing all levels of the training game, listeners performed two experimental tasks. In the first task listeners were asked to identify a voice from among all 15 candidate voices by indicating its *f*0 and FF-scaling level. This task will be referred to as the two-factors task. Listeners identified each voice three times, for a total of 45 trials per participant for this task. The two-factors task should give the best indication of the ability of listeners to separate FF-scaling and *f*0 information, and to report each independently.

For the second task, listeners were again asked to identify stimuli from among all candidate voices; however, for this task listeners only had to indicate stimulus FF-scaling level and ignore *f*0 (this will be referred to as the FF-only task). For this task, only the middle row of response buttons were visible to the listener so that listeners only had the option of reporting FF-scaling. Again, listeners identified each voice three times, for a total of 45 trials per participant for this task. This task was intended to compare the ability of listeners to identify stimulus FF-scaling when listeners are asked to report *f*0 and when they are asked to ignore *f*0. All listeners performed the two-factors task before the FF-scaling only task.

III. RESULTS

The performance of different listeners was expected to vary as a result of two main classes of characteristics. The first of these is the different scaling-factor increments (ΔFF-scale) used to create the synthetic voices. Since larger ΔFF-scales increase the acoustic difference between adjacent FF-scaling levels (i.e., horizontally adjacent voices on the board), it was expected that ΔFF-scale level would affect identification rates, with lower values resulting in worse performance. This is a between-subjects factor in the statistical design and can be dealt with directly as such.

The second class of characteristics expected to affect listener performance is the difference in ability that participants may have had before beginning the training, or the different rates at which participants might learn to independently report the two aspects of voice quality being investigated here. Although no direct measure of these differences is available independently of the experimental results, it was expected that three additional characteristics that relate to listeners' background experience could serve as covariates that reflected these differences in ability.

The first of these covariates is native language, where the performance of native speakers of English might differ from that of non-native speakers. For example, non-native speakers might have more difficulty processing the categorical vowel information and might be operating under a greater cognitive load than native speakers. The second covariate is fluency in a tone language. Seventeen participants were fluent in a tone language. These speakers may have had an advantage in identifying pitch levels or in separating pitch and FF-scaling information relative to speakers without knowledge of a tone language. The final covariate was the number of years of formal musical training a listener had received (including zero for listeners who had received no musical training). In pilot tests of the training program, a listener who was a trained musician performed considerably better than any other listener. It was anticipated that formal musical training might also help listeners learn to separate the *f*0 and FF-scaling information of sounds independently and thus might affect performance. The distribution of these characteristics among listeners in different ΔFF-scale groups is presented in Table II.

A. Identification of voice *f*0 and FF-scaling

1. Performance for the two-factors task

Identification rates were found for correct labeling of *f*0 level, correct labeling of FF-scaling level, and correct absolute identification (where both factors were correctly labeled), individually for each participant ($n = 71$). Performance was high overall with an average of 79.4% *f*0 identifications [min = 31%, max = 100%, standard deviation (sd) = 15.3%], 40.1% correct FF-scaling identifications (min = 15.5%, max = 71%, sd = 12%), and 33.6% correct absolute identifications of both characteristics simultaneously (min = 6.7%, max = 71%, sd = 13.7%). All three

TABLE II. Distribution of some listener characteristics among different ΔFF-scale groups.

ΔFF-scale	7	8	9	10
Total listeners	18	18	18	17
English native speakers	17	14	15	15
Fluent in a tone language	5	4	4	4
Musically trained	7	9	8	6

mean values were considerably higher than what would be expected given chance performance (33%, 20%, and 6.7%, respectively). There was a moderate positive correlation between correct identification rates for *f*0 and FF-scaling within-listeners; listeners who identified *f*0 at a higher rate also identified FF-scaling at a higher rate [$r = 0.44$, $t(69) = 4.1$, $p = 0.0001$].

Listener performance was expected to be affected by the between-subjects factor ΔFF-scale. In addition, the covariates reflecting listeners' background experience were also expected to influence performance level. In order to test which of these characteristics had a significant effect on performance on the two-factors task, a regression analysis was carried out on the within-participant, correct absolute-identification rates. The predictor variables were the between-subjects factor ΔFF-scale (7%, 8%, 9%, 10%), the binary indicator variables native language (English vs non-English), tone language fluency (fluent vs not fluent), and the level of musical instruction, coded as a continuous covariate (in number of years of instruction, including zero for listeners who had received no instruction).

None of the effects reached significance, except the effect of musical training [$F(1,64) = 16.8$, $p = 0.0001$]. Surprisingly, the main effect for ΔFF-scale did not even approach significance [$F(3,64) = 1.4$, $p = 0.25$]. Thus, listeners in the 7% ΔFF-scale group scored about as well as those in the 10% ΔFF-scale group, 37% and 35% correct absolute identifications, respectively. A parallel analysis of variance was carried out on the marginal correct identification rates for voice *f*0 and FF-scaling. These analyses revealed a similar pattern of results with the only significant main effect being for musical training for correct identification of *f*0 [$F(1,64) = 17.8$, $p < 0.0001$] and FF-scaling [$F(1,64) = 9.9$, $p = 0.0025$].

2. Performance for the FF-scaling only task

Since only information regarding FF-scaling estimates was collected for the FF-scaling only task, all references made to correct identification rates refer to FF-scaling identification alone. Once again, correct identification rates were found individually for each participant ($n = 71$). Performance was high overall, with an average correct FF-scaling identification rate of 40.6% (min = 13.3%, max = 64%, sd = 11.8%), which is very close to the 40.1% correct FF-scaling identification rate for the two-factor task.

A regression analysis was carried out in which FF-scaling identification rate was the dependent variable. Once again, the predictor variables were the between-subjects factor ΔFF-scale (7%, 8%, 9%, 10%), the binary indicator variables native language (English vs non-English), tone language fluency (fluent vs not fluent), and the level of musical instruction, coded as a continuous covariate (in number of years of instruction). The same pattern of effects was found as in two-factor task, with only musical training [$F(1,64) = 9.5$, $p = 0.0030$] being a significant predictor of participant performance.

Finally, in order to see if a listener's ability to identify voice FF-scaling was affected by whether they were also asked to report voice *f*0, a *t*-test was carried out on the individual, within-participant difference in FF-scaling identification across the two tasks. The mean within-listener difference in performance between the two task was 0.5%, a difference that did not reach significance [$t(70) = 0.44$, $p = 0.66$]. This indicates that voice FF-scaling estimation is similar in cases where listeners are asked to report voice *f*0 and in cases where they are asked to disregard it.

B. Information used in FF-scaling estimation

The FF-scaling indicated by the listener in response to a trial will be referred to as judged FF-scaling, as opposed to the veridical stimulus FF-scaling level present in each stimulus. Judged FF-scaling is expected to correlate strongly with the listener-internal pFF-scaling perceptual dimension. Consequently, the most important determiner of judged FF-scaling was expected to be stimulus FF-scaling. If listeners were performing this task using only information from the FFs of a vowel pair to determine the FF-scaling of the voice that produced them, stimulus FF-scaling would be the only significant predictor of judged FF-scaling, with no role for stimulus *f*0. On the other hand, a significant main effect for *f*0 may indicate a process of FF-scaling estimation such as Method 6 of the sliding template model (Nearey and Assmann, 2007) where *f*0 may bias FF-scaling estimates. We know of no theory that would predict a significant interaction between *f*0 and FF-scaling in the determination of FF-scaling estimates.

The relationship between judged FF-scaling and stimulus *f*0 and FF-scaling was investigated using ordinal logistic regression. Models of this kind allow one to investigate the classification of stimuli into a sequence of discrete, ordinal categories based on a given number of explanatory variables. In this case, the dependent variable was the judged FF-scaling provided by the listener for each trial. Judged FF-scaling steps were coded as one through five, where higher numbers indicated higher FF-scaling ratings (and higher average FFs for a voice). Stimulus FF-scaling was coded as a centered covariate, while stimulus *f*0 steps were coded using dummy variables, where the lowest *f*0 step acted as the reference group. This coding allows for a linear relationship between stimulus and judged FF-scalings, as well as for stimulus *f*0 levels to result in shifts in judged FF-scaling. The interaction between these two terms allows for the possibility that stimulus FF-scaling had a different linear relationship with judged FF-scaling at different levels of stimulus *f*0.

A model was fit to the data collected for each participant independently, and this was carried out separately for the data from each of the two tasks performed (two-factors task and FF-scaling only task). Significance testing was then carried out on the coefficients found for each listener, for each task, to investigate the effects of each predictor on judged FF-scaling (Gumpertz and Pantula, 1989).

For the two-factors task, stimulus FF-scaling was a highly significant predictor of judged FF-scaling [$F(1,70) = 77.9$, $p < 0.0001$]. As expected, there was a positive relationship between stimulus FF-scaling and judged FF-scaling. The

main effect for $f0$ did not approach significance [$F(2,69) = 0.38$, $p = 0.68$]. However, the interaction between stimulus $f0$ and stimulus FF-scaling was significant [$F(2,69) = 8.79$, $p = 0.0004$].

The interaction between stimulus $f0$ and FF-scaling may be decomposed by stimulus $f0$ level. Since the lowest $f0$ step was used as the reference group, these interactions indicate whether the linear relationship between stimulus and judged FF-scaling differed significantly at the second or third $f0$ steps relative to the relationship observed for the lowest $f0$ step. When considered in this way, only the interaction between the second, intermediate $f0$ level and stimulus FF-scaling reaches significance [$t(70) = -3.08$, $p = 0.0029$]. The interaction is negative, resulting in a decrease in the slope relating stimulus FF-scaling to judged FF-scaling. Since the dependent variable representing stimulus FF-scaling was centered, the decrease in slope indicates that responses tended to gravitate toward the middle of the FF-scaling response space for the middle $f0$ level more so than for the high and low $f0$ levels.

For the FF-scaling only task, there was a very strong positive relationship between stimulus FF-scaling and judged FF-scaling [$F(1,70) = 55.8$, $p < 0.0001$]. Unlike for the two-factors task, stimulus $f0$ [$F(2,69) = 16$, $p < 0.0001$] had a significant (main) effect on judged FF-scaling. The effect of each of the stimulus $f0$ levels on judged FF-scaling was positive, indicating that higher stimulus $f0$'s were associated with higher judged FF-scalings. The interaction between stimulus FF-scaling and stimulus $f0$ was also significant [$F(2,69) = 12.7$, $p < 0.0001$]. When decomposed by stimulus $f0$ level, this interaction showed a similar pattern as that observed for the two-factors task in that only the interaction between the second $f0$ level and stimulus FF-scaling reached significance [$t(70) = 2.66$, $p = 0.0096$]. Once again, this interaction was negative indicating a decrease in the slope relating stimulus FF-scaling to judged FF-scaling.

The significant effects for stimulus $f0$ in both models described above indicate that stimulus $f0$ does have an effect on judged FF-scaling. In order to get a rough estimate of the magnitude of these effects, two linear models were fit to the pooled data across all participants. This process was carried out independently for the results from the two-factors task, and those from the FF-scaling only task. These models treated the response variable, judged FF-scaling, as a continuous variable. The independent variables were coded in the same manner as for the models outlined above. Table III presents the sum of squares and the percent variance explained by each of the explanatory variables for each of these models.

It is clear from the proportion of variance explained by stimulus FF-scaling that judged FF-scaling is most strongly determined by stimulus FF-scaling. In both the two-factors task and the FF-scaling only task, stimulus $f0$ and the interaction between stimulus $f0$ and stimulus FF-scaling explain only a very small amount (0.1% to 3.9%) of the overall variance in judged FF-scaling. These results indicate that the significant effect of stimulus $f0$ on judged FF-scaling, as well as the significant interaction between stimulus $f0$ and stimulus FF-scaling, indicate a small but consistent effect.

TABLE III. Sum of squares and percent of variance explained of judged FF-scaling explained by stimulus FF-scaling (FF-S), stimulus $f0$ ($f0$), and the interaction of the two.

Two-factors task			
Term	df	Sum of squares	% Variance explained
FF-S	1	2244.2	35.6
$f0$	2	32.3	0.5
FF-S × $f0$	2	7.6	0.1
Residual	—	4023.2	63.8
FF-scaling only task			
Term	df	Sum of squares	% Variance explained
FF-S	1	1980.4	31.6
$f0$	2	242.2	3.9
FF-S × $f0$	2	8.9	0.1
Residual	—	4034.3	64.4

IV. DISCUSSION

The motivation behind this experiment was to investigate the extent to which listeners can learn to distinguish and identify voices that vary in average $f0$ and FF-scaling. Results indicate that listeners are able to report voice FF-scaling with reasonable accuracy after only a short training session. Performance was much higher than chance in both the two-factor task and the FF-scaling only task, for absolute identifications of voice FF-scaling and $f0$ where applicable. The high rate at which listeners are able to absolutely identify voice FF-scaling is noteworthy given that the ΔFF-scales used in this experiment (7%–10%) are not much higher than the just noticeable difference in FF-scaling, which has been estimated to be between 4%–8% (Smith *et al.*, 2005; Ives *et al.*, 2005). Furthermore, listeners are able to report voice FF-scaling with the same level of accuracy whether they are asked to report voice $f0$ or to disregard it.

In addition to the high rate at which listeners correctly identified stimulus FF-scaling, their errors tended to cluster around the correct stimulus FF-scaling. Overall, in 65% of errors committed across both tasks, listeners were only off by a single FF-scaling step. In the two-factors task, listeners erred in identifying stimulus FF-scaling by a single step in 39.7% of trials. Combined with the 40.1% of cases in which they correctly identified voice FF-scaling, this means that in 79.8% of trials listeners were either correct or off by a single step. In the FF-scaling only task, they were within one FF-scaling step in 78.8% of cases. By chance alone, listeners would be expected to respond within one step of correct in 52% of cases, meaning that they responded within one step roughly 53% [i.e., (79 – 52)/52] more than expected. These near-miss error patterns suggest that the listener-internal mappings of the stimulus voices are arrayed in a two-dimensional space corresponding closely to $f0$ and FF-scaling. These results all support the notion that there exists a perceptual quality, such as pFF-scaling, which is closely aligned with FF-scaling.

The ability listeners have demonstrated in reporting voice FF-scaling suggests that the experiment reported here could easily be extended to investigate the relationship

between apparent speaker gender and pFF-scaling by instructing listeners that the speaker was of a particular gender on a given trial. A methodology of this kind could be used to investigate the results presented in Johnson *et al.* (1999) and Glidden and Assmann (2004) where changing listener expectations regarding speaker gender affected perceived vowel quality. If a trained listener systematically over- or underestimated stimulus FF-scaling based on apparent speaker gender, it would serve as good evidence that apparent speaker gender affects perceived vowel quality by affecting pFF-scaling estimates based on gender stereotypes.

Another possibility is the use of this training experiment in conjunction with experiments such as those described in Johnson (1990), Johnson *et al.* (1999), and Barreda and Nearey (2012a), in which the relationship between apparent speaker characteristics and perceived vowel quality was investigated. In those experiments, stimulus vowels varied along a limited number of FF dimensions (either $F1$ or $F1$ and $F2$) rather than along all FFs simultaneously, which is the case when they vary in terms of FF-scaling. For example, in Barreda and Nearey (2012a) listeners were presented with vowels that varied along an $F1$-$F2$ continuum, and these were presented with several different $f0$ and higher formant conditions. Apparent speaker size and gender judgments were collected in order to control for estimates of pFF-scaling, and the association between these characteristics and perceived vowel quality was investigated.

However, the results of the experiment presented here suggest that it is possible to ask trained listeners to report speaker FF-scaling directly. For example, given a certain point along the $F1$-$F2$ continuum, we might expect that listeners would respond to changes in the higher formants by indicating different judged FF-scaling levels. Furthermore, given a point along the $F1$-$F2$ continuum, pFF-scaling may co-vary with apparent speaker gender, and perceived vowel quality. Using a methodology of this kind, the relationship between pFF-scaling, apparent speaker characteristics and perceived vowel quality could be investigated more directly.

Barreda and Nearey (2012b) present preliminary results of a study using just this methodology. A replication of Barreda and Nearey (2012a) was carried out in which FF-scaling judgments, as well as speaker gender and vowel quality judgments, were collected from trained listeners. The results indicate that there is a significant relationship between listener FF-scaling responses and reported vowel quality for vowels which had been low-pass filtered above $F3$.[6]

Although listeners are able to report stimulus FF-scaling accurately, some results suggest that the determination of pFF-scaling interacts with the identification of stimulus $f0$ in a complicated manner that warrants further investigation. Correct identification of stimulus $f0$ was associated with higher correct identification of FF-scaling both between-participants (as reported in Sec. III A 1) and within participants; of the 46 listeners who made at least five $f0$ identification errors, FF-scaling identification rates were 6.3% higher when they identified $f0$ correctly relative to cases in which they did not [$t(45) = 2.91$, $p = 0.0056$].

Furthermore, a significant negative correlation was found in errors of $f0$ and FF-scaling identification. A number may be assigned to judged $f0$ and FF-scaling that indicates the difference between these judgments and the veridical stimulus properties. So, for example, zero would indicate a correct identification while negative integers would indicate underestimations and positive numbers would indicate overestimations. For the 46 listeners who made at least 5 $f0$ identification errors, the average within-participant Spearman's correlation coefficient between $f0$ and FF-scaling identification errors was -0.17 [$t(45) = -5.88$, $p < 0.0001$] indicating that FF-scaling overestimations were associated with $f0$ underestimations and vice versa.

The results presented in Sec. III B indicate that stimulus $f0$ has a weak effect on judged FF-scaling, and that this effect can vary for particular combinations of $f0$ and FF-scaling. Furthermore these relationships may vary based on the specific task at hand. For example, in the two-factors task, there was no significant main effect for stimulus $f0$ on judged FF-scaling, while for the FF-scaling only task the main effect for stimulus $f0$ was significant. This may indicate that $f0$ has more of an effect on judged FF-scaling when listeners do not have to explicitly report it, relative to situations in which they do have to report it.

The main effect of $f0$ on judged FF-scaling was positive in cases where it was significant. This is not surprising given the natural co-variation of $f0$ and FF-scaling, where higher $f0$s are associated with higher FF-scalings, and the fact that listeners have demonstrated a sensitivity to this covariation (Assmann and Nearey, 2007, 2008). However, we do not have ready explanations for the interaction patterns observed across the two tasks. In both cases, the linear relationship between stimulus and judged FF-scaling differs for the middle $f0$ step relative to the high and low $f0$ steps, and this difference manifested itself as a decrease in the positive relationship between the two variables, resulting in a compression toward the middle of the response space.

These results suggest that $f0$ may play a role in the determination of pFF-scaling, and that this may not be determined solely on the basis of the FFs of a vowel sound. An effect for $f0$ on pFF-scaling is predicted by Method 6 of the sliding template model of Nearey and Assmann (2007), where they suggest that pFF-scaling (which they refer to as ψ) is determined party on the basis of $f0$. However, this model would only predict linear shifts in pFF-scaling based on $f0$, and not a complicated pattern of interactions. This model also has no way to explain the negative correlation of errors observed, not the varying effect of $f0$ based on task type.

The significant and complicated effect of stimulus $f0$ on judged FF-scaling casts doubt on the theories put forth by Irino and Patterson (2002), Smith *et al.* (2005), and Turner *et al.* (2006). These researchers claim that the peripheral auditory system performs transformations on speech sounds that automatically segregate information related to vocal-tract configuration from information related to FF-scaling, and that human listeners have direct access to FF-scaling information resulting from this processing. If this were the case, there is no clear reason why $f0$ should significantly

influence directly reported FF-scaling judgments, or for this influence to vary based on task. Although transforms such as those suggested by these authors may still occur, a transformation which segregates information regarding voice FF-scaling, only to recombine it with *f*0 information before the listener can access it would not be of much use to listeners.

Some characteristics of the experimental design make it unsuitable questions regarding the processes that underlie the construction of a pFF-scaling dimension, and the manner in which this is influenced by *f*0. This experiment was designed to investigate whether listeners are able to identify voices on the basis of their FF-scaling, and whether it would be feasible to collect FF-scaling estimates from listeners in perceptual experiments.

First, the sampling of the *f*0 dimension was deliberately sparse, and many listeners committed very few, or no *f*0 identification errors at all. For example, 35 of 71 listeners made less than 5 *f*0 identification errors out of a total of 45 trials for the two-factors task. We did not want to present too complex or frustrating a task to listeners until we were certain they could reliably respond to FF-scaling differences in voices. Second, the sampling of the FF-scaling dimension was intended to replicate the stimulus design of experiments that might involve the collection of FF-scaling estimates rather than to investigate the process of FF-scaling estimation as a continuous dimension. Finally, the limited number of trials carried out for each of the two tasks makes it difficult to analyze these processes in great detail. However, it is important to note that the effect of *f*0 and the correlation of errors were detectable despite these shortcomings, which suggests that these are important considerations in the construction of a pFF-scaling dimension.

In the future, experiments with stimuli that more densely sample the *f*0 × FF-scaling space, and which feature a higher number of trials will need to be carried out to investigate more specific questions regarding the processes involved in *f*0 and FF-scaling estimation. Of particular interest to the field of speech perception is the way in which these two processes may cooperate and the ways in which this cooperation may interact with the estimation of apparent speaker characteristics and the determination of vowel quality.

V. CONCLUSION

The experiment outlined here involved a training method in which listeners learned to report voice FF-scaling. Although listeners have previously demonstrated a sensitivity to changes in voice FF-scaling independently of *f*0, the average listener may not have a ready label for the acoustic characteristic associated with the average FFs produced by a voice. Results indicate that listeners are able to provide FF-scaling judgments with relative ease and consistency, and that these estimates are most strongly determined by the FFs of a stimulus, with only weak effects for stimulus *f*0. This may be contrasted with apparent speaker characteristics such as apparent speaker size and gender, which are most strongly determined by the *f*0 of a vowel, with a weaker effect for the FFs (Gelfer and Mikos, 2005; Hillenbrand and Clark, 2009).

The results presented here suggest that it is feasible to collect FF-scaling estimates from listeners in further experiments which seek to investigate the process of FF-scaling estimation, or the role of FF-scaling estimation in speech perception. Furthermore, they suggest that there exists a perceptual dimension closely aligned with FF-scaling (i.e., pFF-scaling), and that this perceptual dimension may be influenced to some extent by *f*0 in a complicated manner that is not explained by any theory we are aware of. Given the potential importance of FF-scaling, and its perceptual counterpart pFF-scaling, for vowel perception and the determination of apparent speaker characteristics, these issues warrant further investigation.

APPENDIX

It has been suggested that non-uniformities in the vocal tracts of speakers of different sizes might result in the non-uniform scaling of speech sounds between adult males and other speakers (Fant, 1975). Fant suggested that such non-uniformities were due to the relatively longer pharynx-to-mouth ratios of adult males. However, no clear demonstration either of the statistical reliability of systematic non-uniformities nor of the perceptual relevance of any such non-uniformities to listeners' identification performance exist in the literature.

Turner *et al.* (2009) review difficulties with this hypothesis. In particular, they present a re-examination of the physiological data reported by Fitch and Giedd (1999) and find that although the oral-pharyngeal cavity ratios vary continuously in relation to speaker size, and not simply on the basis of speaker gender, there is no evidence that these differences manifest themselves as differences in produced formant patterns. They conclude that "the anatomical distinction between the oral and pharyngeal divisions of the vocal tract is immaterial to the acoustic result of speech production. For a given vowel, the tongue constriction is simply positioned where it produces the appropriate ratio of front-cavity length to back-cavity length, independent of the location of the oral-pharyngeal junction" (p. 2379). They also state that "speakers adjust the shape of the vocal tract as they grow to maintain a specific pattern of formant frequencies for individual vowels" (p. 2374). Basically, despite differences in anatomy from person to person, speakers strive to produce vowels which differ by a single parameter (i.e., FF-scaling) from the same vowel when produced by other speakers of their language, even if this entails slight modifications to articulatory gestures as a speaker ages.

We do not intend to suggest that vowels vary within-category, between-speakers, solely on the basis of FF-scaling in a deterministic manner. Rather, our position is that, all other things being equal, vowels from speakers of the same dialect with varying vocal-tract lengths differ in terms of this parameter plus statistical noise. This noise may result from idiosyncratic differences in articulation or speaker anatomy, or it may be a result of the particular situation in which the speech was produced (e.g., clear vs casual speech). The left panel of Fig. 3 shows the classic Peterson and Barney (1952) vowel data. A visual inspection of Fig. 3

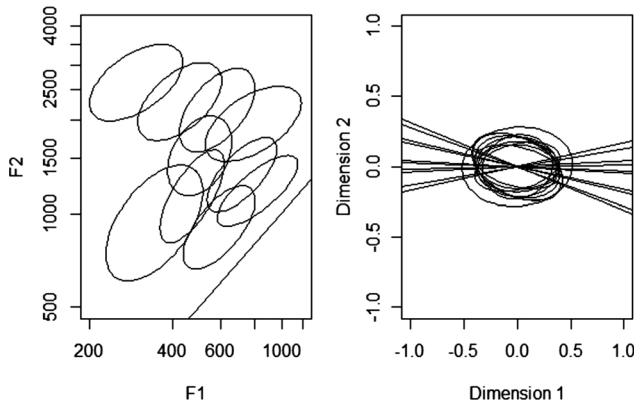

FIG. 3. In the left panel, ellipses enclosing two standard deviations of the Peterson and Barney (1952) vowels are presented. The line is a line parallel to $\ln F1 = \ln F2$. In the right panel, all formant frequencies have been log-transformed and centered within-category so that vowel-category means are at the origin. All points have been rotated 45 deg clockwise so that the $\ln F1 = \ln F2$ line is now parallel to the x-axis (Dimension 1), while Dimension 2 represents the orthogonal axis. The lines indicate the major axes of the vowel-category ellipses, and they all vary around the $\ln F1 = \ln F2$ axis (Dimension 1).

clearly shows that the major axes of the ellipses are aligned with the $F1 = F2$ line in a log-space (henceforth, $\ln F1 = \ln F2$), also indicated on the figure. Variation along the $\ln F1 = \ln F2$ indicates equal logarithmic increases to both $F1$ and $F2$, and is consistent with variation according to a single multiplicative parameter.

To investigate the extent of variation along the $\ln F1 = \ln F2$ axis, the following analysis was carried out.[7] Formant frequencies were log transformed, and centered according to vowel-category so that all category means were located at the origin. After this, all points were rotated by 45 deg in a clockwise direction. The result of this is presented in the right panel of Fig. 3. As a result of these transformations, the x-axis now represents a line parallel to $\ln F1 = \ln F2$ and variation along this axis represents variation within vowel-category, between-speakers, that results from uniform logarithmic increases to $F1$ and $F2$ (i.e., by a single multiplicative parameter). This analysis revealed that 80.6% of variation between-speaker falls along the $\ln F1 = \ln F2$ axis. The same analysis carried out on the vowel data of Hillenbrand *et al.* (1995) revealed that 79.6% of variation in FFs between speakers falls along the $\ln F1 = \ln F2$ axis for that data set. These results are consistent with the hypothesis that variation in FFs within vowel-category, between-speakers, is largely according to a single multiplicative parameter.

[1]Lack of significance could be in part due to the reduced power of tests based on a small number of observations compared to the full sample. This is at least partly due to the restricted ranges used when considering correlations between acoustic characteristics of speech and the physical qualities of the speaker only for a restricted class of speakers. By restricting the range of a predictor when the error in the response variable remains constant, the correlation between two variables will become weaker (Bland and Altman, 2011; Sackett and Yang, 2000). In the most extreme example, the correlation between the acoustic properties of voices and the heights of men who are all the same height will necessarily be zero.

[2]As far as we have been able to determine, this perceptual property has no specific name in either psychophysical or musical terminology, although it appears to bear some relation to some subdivisions of the German Fach system of classification of operatic voices. Such a perceptual property might correspond to the scale-dimension of what Patterson and colleagues propose is a Mellin(-like) transform performed by the peripheral auditory system that segregates information related to vocal-tract length from information relation to vocal-tract configuration. In Sec. I B, we suggest that pFF-scaling might be a kind of derived perceptual property, which is determined when a listener establishes a speaker-dependent frame of reference. The location of that frame of reference is indexed by a single scalar value, analogous to ψ in Nearey and Assmann's (2007) sliding template model, and the parameter a seen in Eq. (1) presented in Turner *et al.* (2009, p. 2377).

[3]The classifier was also used to predict vowel category for the Peterson and Barney (1952) data set. However, unlike for the other data sets, these predictions did not use vowel formant onset and offset information, which has been demonstrated to significantly improve vowel identification accuracy (see, e.g., Hillenbrand *et al.*, 1995). Since the human listeners who classified vowels in Peterson and Barney undoubtedly had access to this information, the comparison of performance between the two is not appropriate. The lower identification for the Katz and Assmann (2001) vowels relative to the Hillenbrand *et al.* (1995) vowels is very likely due to the fact that in Assmann and Katz, 60% of the vowel tokens came from children between the ages of 3 and 7, and exhibited a higher degree of variability in formant frequencies than the vowels produced by adult speakers.

[4]A positive relationship was expected between perceived vowel quality and apparent speaker size, and 14 of 19 participants exhibited a positive relationship between the two variables. This corresponds to a one-tailed p-value of 0.0318 using a non-parametric sign test. However, a t-test of the same partial correlation finds that they are not significantly different from zero ($p = 0.3027$).

[5]In our analysis we will assume listener's judgments are really separated into these two components at the time of choice. However, even if listeners were instead memorizing a discrete set of individual voices, the systematic correspondence of their choices to the FF-scaling and $f0$ dimensions would at least provide evidence that the "perceptual speaker space" is organized in a way that includes a subspace that is effectively near projection of these two dimensions.

[6]However, to our surprise, this was not the case for vowels with more higher formants. The presence or absence of higher formants had a complicated relationship with apparent speaker gender and reported pFF-scaling. This may have resulted in a weakening of the relationship between reported pFF-scaling and reported vowel quality.

[7]This analysis is similar to one presented in Turner *et al.* (2009). However, that analysis was based on formant wavelengths rather than log-transformed formant-frequencies, which may result in unstable variances. Furthermore, Turner *et al.* allowed for a specific principal component for each vowel-category ellipse, rather than calculating variation strictly along the axis corresponding to changes in FFs by a single parameter. Allowing for a category-specific slope, and allowing these to vary away from parallelism to the $\ln F1 = \ln F2$ line makes that analysis incompatible with a strict uniform scaling hypothesis.

Ainsworth, W. (**1975**). "Intrinsic and extrinsic factors in vowel judgments," in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. Tatham (Academic, London), pp. 103–113.

Assmann, P. F., and Katz, W. F. (**2000**). "Time-varying spectral change in the vowels of children and adults," J. Acoust. Soc. Am. **108**(4), 1856–1866.

Assmann, P. F., and Nearey, T. M. (**2007**). "Relationship between fundamental and formant frequencies in voice preference," J. Acoust. Soc. Am. **122**(2), EL35–EL43.

Assmann, P. F., and Nearey, T. M. (**2008**). "Identification of frequency-shifted vowels," J. Acoust. Soc. Am. **124**(5), 3203–3212.

Barreda, S., and Nearey, T. M.. (**2012a**). "The direct and indirect roles of fundamental frequency in vowel perception," J. Acoust. Soc. Am. **131**, 466–477.

Barreda, S., and Nearey, T. (**2012b**). "The association between speaker-dependent formant space estimates and perceived vowel quality," Can. Acoust. **40**, 12–13.

Bland, J. M., and Altman, D. G. (**2011**). "Correlation in restricted ranges of data," BMJ **342**, d556.

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

Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, The Hague), pp. 107–138.

Fant, G. (**1975**). "Non-uniform vowel normalization," Speech Transm. Lab. Q. Prog. Status Rep. **2–3**, 1–19.
Fitch, W. T. (**1994**). "Vocal tract length perception and the evolution of language," Doctoral dissertation, Brown University.
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.
Gelfer, M. P., and Mikos, V. A. (**2005**). "The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels," J. Voice **19**, 544–554.
Glidden, C., and Assmann, P. F. (**2004**). "Effects of visual gender and frequency shifts on vowel category judgments," ARLO **5**, 132–138.
Gonzalez, J. (**2004**). "Formant frequencies and body size of speaker: A weak relationship in adult humans," J. Phonetics **32**, 277–287.
Gumpertz, M., and Pantula, S. G. (**1989**). "A simple approach to inference in random coefficient models," Am. Stat. **43**(4), 203–210.
Hillenbrand, J. M., and Clark, M. J. (**2009**). "The role of F0 and formant frequencies in distinguishing the voices of men and women," Attention, Percept., Psychophys. **71**, 1150–1166.
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.
Hollien, H., Green, R., and Massey, K. (**1994**). "Longitudinal research on adolescent voice change in males," J. Acoust. Soc. Am. **96**, 2646–2653.
Honorof, D. N., and Whalen, D. H. (**2005**). "Perception of pitch location within a speaker's F0 range," J. Acoust. Soc. Am. **117**(4), 2193–2200.
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.
Ives, D. T., Smith, D. R. R., and Patterson, R. D. (**2005**). "Discrimination of speaker size from syllable phrases," J. Acoust. Soc. Am. **118**, 3816–3822.
Johnson, K. (**1990**). "The role of perceived speaker identity in *f*0 normalization of vowels," J. Acoust. Soc. Am. **88**, 642–654.
Johnson, K. (**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.
Johnson, K., Strand, E. A., and D'Imperio, M. (**1999**). "Auditory-visual integration of talker gender in vowel perception," J. Phonetics **27**, 359–384.
Joos, M. (**1948**). "Acoustic phonetics," Language **24**, 1–136.
Katz, W. F., and Assmann, P. F. (**2001**). "Identification of children's and adults' vowels: Intrinsic fundamental frequency, fundamental frequency dynamics, and presence of voicing," J. Phonetics **29**, 23–51.
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**(2), 820–857.
Labov, W., Ash, B. S., and Boberg, C. (**2006**). *The Atlas of North American English: Phonetics, Phonology and Sound Change* (Walter de Gruyter, Boston).
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.
Lee, S., Potamianos, S., and Narayanan, S. (**1999**). "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," J. Acoust. Soc. Am. **105**, 1455–1468.
Miller, R. L. (**1953**). "Auditory tests with synthetic vowels," J. Acoust. Soc. Am. **25**, 114–121.
Nearey, T. M. (**1978**). "Phonetic feature systems for vowels," Ph.D. thesis, Indiana University Linguistics Club.
Nearey, T. M. (**1989**). "Static, dynamic, and relational properties in vowel perception," J. Acoust. Soc. Am. **85**, 2088–2113.
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. Beddor, and M. Ohala (Oxford University Press, Oxford), pp. 246–269.
Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.
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.
Sackett, P. R., and Yang, H. (**2000**). "Correction for range restriction: An expanded typology," J. Appl. Psychol. **85**(1), 112.
Slawson, A. W. (**1968**). "Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency," J. Acoust. Soc. Am., **43**(1), 87–101.
Smith, D. R. R., and Patterson, R. D. (**2005**). "The interaction of glottal-pulse rate and vocal-tract length in judgments 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, T. (**2005**). "The processing and perception of size information in speech sounds," J. Acoust. Soc. Am. **117**, 305–318.
Turner, R. E., Al-Hames, M. A., Smith, D. R. R., Kawahara, H., Irino, T., and Patterson, R. D. (**2006**). "Vowel normalisation: Time-domain processing of the internal dynamics of speech," in *Dynamics of Speech Production and Perception*, edited by P. Divenyi (IOS, Amsterdam), pp. 153–170.
Turner, R. E., Walters, T. C., Monaghan, J. J. M., and Patterson, R. D. (**2009**). "A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data," J. Acoust. Soc. Am. **125**(4), 2374–2386.
van Dommelen, W. A., and Moxness, B. H. (**1995**). "Acoustic parameters in speaker height and weight identification: Sex-specific behaviour," Lang. Speech **38**, 267–287.