Perception & Psychophysics
1997, 59 (5), 729-738

Intrinsic *f*0 differences in spoken and sung vowels and their perception by listeners

CAROL A. FOWLER
Haskins Laboratories, New Haven, Connecticut
University of Connecticut, Storrs, Connecticut
and Yale University, New Haven, Connecticut

and

JULIE M. BROWN
Haskins Laboratories, New Haven, Connecticut
and University of Connecticut, Storrs, Connecticut

1044

We explore how listeners perceive distinct pieces of phonetic information that are conveyed in parallel by the fundamental frequency (*f*0) contour of spoken and sung vowels. In a first experiment, we measured differences in *f*0 of /i/ and /a/ vowels spoken and sung by unselected undergraduate participants. Differences in "intrinsic *f*0" (with *f*0 of /i/ higher than of /a/) were present in spoken and sung vowels; however, differences in sung vowels were smaller than those in spoken vowels. Four experiments tested a hypothesis that listeners would not hear the intrinsic *f*0 differences as differences in pitch on the vowel, because they provide information, instead, for production of a closed or open vowel. The experiments provide clear evidence of "parsing" of intrinsic *f*0 from the *f*0 that contributes to perceived vowel pitch. However, only some conditions led to an estimate of the magnitude of parsing that closely matched the magnitude of produced intrinsic *f*0 differences.

Our research explores a characteristic of speech perception that we have called "parsing" (e.g., Fowler, 1996). Parsing occurs when different phonetic components of an utterance have converging effects on common acoustic dimensions. Convergences of this sort can be consequences of coproduction of phonetic segments with one another or of phonetic segments with speech prosody. Due to vowel–vowel coproduction and vowel–consonant coproduction, for example, the acoustic spectrum at most points in time is influenced by two or more phonetic segments. The listener's parsing of context-sensitive acoustic structure is revealed jointly by two kinds of findings.

First, and focusing on segment–segment coarticulation to begin with, we find that listeners use acoustic information caused by a coarticulating consonant or vowel as information for the coarticulating segment. For example, listeners use coarticulatory information in one vowel that has been caused by a coarticulating following vowel as information for the following vowel (Fowler & Smith, 1986; Martin & Bunnell, 1981). Accordingly, in the nonsense disyllables /bəbi/ and /bəba/, listeners use coarticulatory evidence of raising and fronting during /ə/ in the first disyllable and of lowering and backing during /ə/ in the second disyllable as information for the identity of the disyllable-final /i/ or /a/ (Fowler & Smith, 1986). Compatibly, they use information in a consonant caused by a vowel as information for the vowel (Fowler, 1984) and vice versa (Whalen, 1984).

Listeners might be able to use coarticulatory information in these context-sensitive acoustic signals for either of two reasons. First, they might hear the context sensitivity as such. That is, the acoustic signal for segment *a*, which has been affected by anticipatory coarticulation from a segment *b*, might sound like a_b, and listeners might use the *b*-like character of *a* to predict that the forthcoming segment is *b*. Alternatively, as suggested by the motor theory (see, e.g., Liberman & Mattingly, 1985) or the theory of direct perception (see, e.g., Fowler, 1986), listeners may partition the acoustic signal along phonetic-gestural lines. In these theories, primitives of spoken words are phonetic gestures of the vocal tract, and listeners perceive gestures as the smallest linguistically significant components of an utterance. A listener who perceives gestures will use the constellation of acoustic consequences of the gestures as information for them. Accordingly, given a_b, listeners will parse the acoustic consequences of anticipating gesture *b* during production of *a* and will use it as information for *b*, not as context sensitivity of *a*.

Evidence suggests that listeners perceive in the latter way. This is the second finding relating to parsing to which we alluded above. For example, in the following set of four trisyllables (Fowler, 1981; Fowler & Smith, 1986), subscripts identify the original phonetic context in which spliced medial /bə/ syllables had been produced; accordingly, below, ${}_{i}b\text{ə}_{i}$ had originally been produced in $i_{i}b\text{ə}_{i}bi$ and ${}_{a}b\text{ə}_{a}$ had been produced in $a_{a}b\text{ə}_{a}ba$:

This research was supported by NICHD Grant HD-01994 to Haskins Laboratories. Correspondence should be addressed to C. A. Fowler, Haskins Laboratories, 270 Crown St., New Haven, CT 06511 (e-mail: fowler@haskins.yale.edu).

A B

$i_ibə_ibi$ -- $a_ibə_iba$ ---------- $i_ibə_ibi$ -- $a_abə_aba$.

Listeners were asked to decide in which pair of trisyllables, A or B, the medial unstressed syllables sounded more alike. If listeners base their judgments solely on the context-sensitive medial syllables, they should judge the medial syllables in Pair A as more alike than those in Pair B, because, in fact, those in A are acoustically identical. However, if listeners ascribe to the flanking /i/ and /a/ gestures the coarticulatory effects that each should have caused in the schwa vowels, then schwas in Pair B should sound identical, whereas those in A should sound different. That is, in both $i_ibə_ibi$ trisyllables, listeners should parse the raising and fronting effects of the flanking /i/s in the domain of the schwa vowels, leaving uncoarticulated schwa as remainder. Likewise in $a_abə_aba$, listeners will parse the lowering and backing effects of /a/ in the domain of schwa, once again leaving uncoarticulated schwa as remainder. However, in $a_ibə_iba$, listeners will parse the lowering and backing effects that the flanking /a/ should have had on the medial vowel, and the residual unstressed vowel will sound high and front. In two experiments (Fowler, 1981; Fowler & Smith, 1986), listeners chose trisyllables in pairs such as B as having the more similar schwas.

In the present research, we explore parsing that may occur when phonetic segments and prosodic structure have converging effects on a common acoustic dimension. Here we focus on vocalic phonetic segments and speech melodies or intonational patterns that have converging effects on fundamental frequency ($f0$).

Other things equal, high vowels are associated with higher $f0$s than are low vowels (see, e.g., Sapir, 1989, for a review). This is called a difference in intrinsic fundamental frequency. The reasons for intrinsic $f0$ are controversial.

Investigations have shown that cricothyroid muscle activity is higher during the production of high vowels than during that of low vowels (Dhyr, 1990; Honda & Fujimura, 1991; Vilkman, Aaltonen, Raimo, Arajärvi, & Oksanen, 1989). The cricothyroid is a laryngeal muscle the activation of which can have several effects, one of which is to raise $f0$. Accordingly, the finding may imply that talkers create the confounding of vowel height and $f0$ intentionally. Raising $f0$ on high vowels, which have low first formants ($F1$s), moves $f0$ closer to $F1$ and may, therefore, enhance their distinctness from low vowels, which have high $F1$s but low $f0$s (Diehl, 1991; Kluender, 1994). However, Whalen and Levitt (1995) have argued that the perceptual enhancement of high vowels caused by a high $f0$ should be negligible. Further, the cricothyroid muscle has consequences other than raising $f0$; possibly one or more of these consequences explains its activity during high vowels.

Some investigators (Honda, 1981; Ohala & Eukel, 1976) have proposed that the increase in $f0$ for high vowels may be due in part to the tongue's pulling on tissues or structures of the larynx. Compatibly, there is evidence that intrinsic $f0$ is exaggerated when speakers produce vowels with a large biteblock holding the jaw in a very open position (Ohala & Eukel, 1987). Given that the biteblock should enhance any tongue pull, Ohala and Eukel's study provides some supportive evidence for a mechanism of this type. Further, one tongue pull account suggests a reason for cricothyroid activity accompanying high vowels. If the tongue pull works to enlarge the space between the thyroid and cricoid cartilages (the "cricothyroid visor"), then, as Vilkman et al. proposed (1989; see also Vilkman, Aaltonen, Laine, & Raimo, 1991), action of the cricothyroid may function to offset that effect of the pull. If so, any increase in $f0$ for high vowels that is due to cricothyroid activity may be an incidental, rather than an intended, consequence of producing high vowels.

Despite the positive evidence for an account of intrinsic $f0$ that invokes tongue pull, there is evidence opposing the account as well. For example, Fischer-Jørgensen (1990) summarized evidence that in comparisons of tense–lax vowel pairs, differences in tongue height and in intrinsic $f0$ do not pattern as they should according to the tongue pull account. Indeed, currently no explanation for intrinsic $f0$ accounts for all of the relevant data satisfactorily (see, e.g., Sapir, 1989; Silverman, 1987, for reviews of the variety of accounts of intrinsic $f0$ and the relevant data).

On two indirect, but relevant grounds, Whalen and colleagues (Whalen & Levitt, 1995; Whalen, Levitt, Hsaio, & Smorodinsky, 1995) suggested that, whatever account of intrinsic $f0$ turns out to be accurate, it is likely to be one in which the effect is an incidental, rather than an intended, characteristic of spoken vowels. First, intrinsic $f0$ is present in the babbling of 6- to 12-month-old infants from English- and French-speaking homes (Whalen et al., 1995). Second, it appears to be universal to languages and not to vary among languages that differ in the sizes of their vowel inventories from 4 to 18 vowels (Whalen & Levitt, 1995). Neither babbling infants nor speakers with four-vowel inventories need to use $f0$ to distinguish their high from their low vowels for listeners—babbling infants because they are not communicating, and speakers of four-vowel inventories because their vowels are spectrally highly distinct already.

We will not attempt to discriminate accounts of intrinsic $f0$ as incidental or intended in any direct way in the research that we report here. However, in a theory in which listeners track phonetic gestures, parsing of $f0$ into a "vowel height" component and an "intonation" or "pitch" component is expected only if the phonetic gesture that underlies production of the high vowel causes the increase in $f0$. Accordingly, evidence of $f0$ parsing favors accounts of intrinsic $f0$ that invoke incidental rather than intended sources.

Across studies, parsing has already been shown for perceived vowel height and intonational accent. Reinholt-Peterson (1986) showed that a vowel ambiguous between the Danish vowels /o:/ and /u:/ was more likely to be perceived as the higher /u:/ when vowels were synthesized with a higher than a lower $f0$. This shows that $f0$ provides information for vowel height that listeners use. On the

other hand, Silverman (1987) showed that when the high vowel /i/ and the low vowel /a/ had identical $f0$ contours in a sentential frame in which each received an intonational accent, /i/ was judged to underlie a less prominent accent than /a/. Apparently, $f0$ that is ascribed to vowel height is parsed from the $f0$ contour, with the remainder heard as the intonational accent. Under some conditions, Silverman found a magnitude of parsing that closely approximated the amount (10–15 Hz) that he estimated from the literature to be characteristic of intrinsic $f0$ differences between high and low vowels.

Our research pursues this finding in two ways. First, we attempt to replicate Silverman's (1987) findings that the amount of $f0$ that listeners ascribe to vowel height tends to match the magnitude of intrinsic $f0$ differences produced by talkers. Second, we extend the observation of parsing to sung vowels. For sung high and low vowels to have the same pitch, or for sung high and low vowels to match the pitch of a tone, must singers produce /i/s that exceed /a/s in sung $f0$ and that exceed tones in frequency?

EXPERIMENT 1

The first experiment was designed in part to provide an estimate of the magnitude of intrinsic $f0$ differences between the high vowel /i/ and the low vowel /a/ in normally spoken, isolated words and when they are sung. In addition, it provides a natural speech source of the resynthesized speech used as perceptual stimuli in Experiments 2–5.

As noted, Silverman (1987) estimated intrinsic $f0$ magnitude (between high and low vowels) to range between 10 and 15 Hz. Intrinsic $f0$ is present in the singing of trained singers, but it is considerably smaller in magnitude than in intrinsic $f0$ in ordinary speech (Grieffenberg & Reinholt-Petersen, 1982; Ternström, Sundberg, & Colldén, 1988). In the present experiment, we obtained estimates of intrinsic $f0$ in sung vowels produced by individuals who were unselected for training in singing.

Method

Subjects. The subjects were native speakers of English who were undergraduates; all reported normal speech and hearing. They received course credit for their participation. Thirteen subjects were run; after data collection and measurement, a tape recording of 5 of the speakers was inadvertently destroyed. We present separate analyses of the data of all 13 subjects, and of those of just the 8 whose recordings remained available.

Procedure. Each subject took part in all three production tasks. The order of tasks was rotated across subjects. For the speech task, the students sat facing a computer monitor that presented printed words one at a time to be read. The words *beady*, *body*, *keyed*, and *cod* were presented 10 times each in random order with a 5-sec interval between them. The students were asked to read the words at a leisurely pace.

In one singing condition (henceforth the "vowel-shifting" condition), a computer monitor cued the subjects to begin singing either /i/ or /a/ at a self-selected pitch. Their instructions were to sing the vowel on a constant pitch; they were to sustain the vowel for approximately 2 sec and then shift to the other vowel, keeping pitch constant, and to hold the second vowel for another 2 sec. They produced 10 /a/–/i/ sequences and 10 /i/–/a/ sequences, randomly ordered.

In a second singing condition (the "tone-matching" condition), on each of 24 trials, the computer cued the vowel, /i/ or /a/, that the subject was to sing. A tone then sounded and continued for 10 sec. The tone was a 150-Hz sine wave for male subjects and a 250-Hz sine wave for females. The subjects were instructed to sing the vowel to match the tone. Once they were satisfied that they were matching the tone, they were asked to hold the vowel for 2 sec. In all conditions, sung vowels were recorded on audiotape.

Measurements. In the spoken words, $f0$ was measured in the vicinity of the /i/ or /a/ vowel midpoint. This was done in either of two ways. It was extracted using an algorithm provided by the MacSpeechLab (GW Instruments) software, or, if the algorithm appeared to provide spurious results, it was extracted by measuring the duration of pitch pulses surrounding the vowel midpoint.

In the first singing condition, each vowel was measured at three locations. For the first vowel in a sequence, the first measurement was made approximately one-half second from the beginning of the vowel; another measurement was made just before the transition to the second vowel; and a measurement was made halfway between these measurement points. For the second vowel, measurements were made just after the transition, at a point approximately one-half second from the end of the vowel and at a point between these two measurements. In the second condition, three measurements were made, evenly separated during the last 2 sec of the sung vowel. Measurements were made in the same way (using MacSpeechLab) as in the speech condition. Because there were no statistically significant differences among the three measured points in either singing condition, for purposes of subsequent analyses, we averaged across the points.

Results and Discussion

Results of all three conditions are shown in Figure 1. The left pair of bars in Figure 1 displays the mean $f0$s for the /i/ and /a/ words averaged across the 13 subjects. The middle set of bars shows the mean $f0$ in the first (vowel-shifting) singing condition, and the right-most pair of bars shows the results of the tone-matching task. The differences in $f0$ shown in the figure also closely represent the $f0$ differences shown by the subset of 8 subjects on whom we have the original recordings.

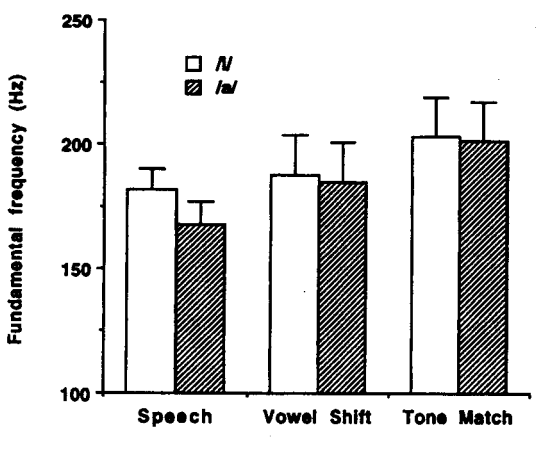

Figure 1. Average $f0$ (Hz), with standard error bars, of /i/ and /a/ vowels in spoken words (left pair of bars), in vowels sung at a self-selected pitch (middle bars), and in vowels sung to match a tone (right bars). Note that error bars reflect between-subjects (including between-sex) variability in $f0$, not the within-subjects variability that determined significance in statistical tests.

Speech condition. In analyses of the speech condition with the factors consonant (/b/, /k/) and vowel (/i/, /a/), only the effect of vowel was significant [all 13 subjects, $F(1,12) = 61.00$, $p < .001$; the subset of 8, $F(1,7) = 34.58$, $p = .0006$]. Analyses performed separately on the data from each subject revealed a significant vowel effect for all 13 speakers. Although the literature suggests that an effect of consonant should have been found (see, e.g., Hombert, 1978), with *f*0s following /k/ higher than /b/, these data did not show a difference; accordingly, Figure 1 presents the data collapsed across the consonant factor.

Averaging over the consonants, the intrinsic *f*0 effect (the difference in average *f*0 for /i/ minus /a/) was 13 Hz in the set of 13 subjects and 14 Hz in the subset of 8 subjects. This leads us to predict that for vowels /i/ and /a/ to sound to listeners as if they are equal in pitch, /i/ should exceed /a/ in *f*0. More specifically, if listeners are accurate parsers, /i/ should exceed /a/ by 13–14 Hz at the point of pitch equality. Experiments 2 and 3 tested those predictions.

Vowel shifting. In the analysis of the first singing condition, in which subjects sang one vowel and then shifted to the other, there was a 3.00-Hz difference in frequency between sung /i/ and /a/ across the 13 subjects [$t(12) = 6.43$, $p < .0001$]. Across the subset of 8 subjects, the difference was 2.63 Hz [$t(7) = 6.25$, $p = .0004$]. In separate analyses on individuals, the effect was significant for 11 of the 13 subjects and marginal for a 12th ($p = .053$).

Tone matching. In the analysis of the tone data of the 13 subjects, there was a 1.85-Hz difference in the frequency of sung /i/ and /a/ in the predicted direction [$t(12) = 3.15$, $p = .008$]. Across the subset of 8 subjects, the difference was 2.34 Hz [$t(7) = 2.89$, $p = .023$]. In separate analyses on individuals, only 5 of the 13 (and 3 of the 8) subjects showed significant /i/ − /a/ differences in this condition. (Male and female subjects, who had matched vowels to different tones reflecting their different average fundamental frequencies, showed nearly the same—approximately 2 Hz—effect sizes.)

With three exceptions, subjects came close to matching the tones. For the 10 successful subjects, the average absolute value of the discrepancy between the sung *f*0 and the frequency of the tone was 3.0 Hz. (For the other 3 singers, the difference averaged 50.8 Hz.) Although the majority of close matchers sang /i/ at a higher frequency than the tone (8 of 10), the majority also sang /a/ at a higher frequency than the tone (7 of 10).

Replicating findings by Grieffenberg and Reinholt-Petersen (1982) and by Ternström et al. (1988), we have found small but reliable differences in the intrinsic *f*0 of high and low sung vowels.

Cross-condition analysis. In an analysis across the three conditions, with the factors condition (spoken, sung with vowel shift, sung to match a tone) and vowel (/i/, /a/), we found significant main effects, but more importantly an interaction between them [$F(2,24) = 11.41$, $p = .0043$, across the 13 subjects; $F(2,14) = 13.28$, $p = .007$, across the subset of 8 subjects]. The interaction reflects the significantly larger intrinsic *f*0 difference in spoken than in sung /i/ and /a/.

This outcome led us to predict less parsing of intrinsic *f*0 from the sung than from the spoken vowels. We predicted that listeners would parse approximately 3 Hz from the sung vowels of the vowel-shifting condition and approximately 2 Hz from the tone-matching condition. These predictions were tested in Experiments 4 and 5.

EXPERIMENT 2

To test whether listeners will parse *f*0 from the intonation contour on which speakers produced the words of Experiment 1, we used resynthesis on a typical speaker's natural productions of the words *keyed* and *cod* from the first experiment. We created several versions of each of the two words that were identical except for their *f*0 contours. Listeners judged which of a *keyed–cod* pair that they had heard on each trial had the higher pitch. We expected them to hear *cod* as higher in pitch when the *f*0 contours on the vowels were identical, and we expected them to require approximately a 13-Hz difference in *f*0 for the pitch of the words to sound equal.

Method

Subjects. Twenty-two introductory psychology students participated in Experiment 2 for course credit. All were native English speakers with normal hearing.

Materials. Recordings of 1 of the male subjects from Experiment 1, whom we will call M1, were used for this experiment. His data were typical in showing considerably larger intrinsic *f*0 differences in the spoken-word than in either singing condition. We used one token each of M1's spoken productions of *cod* and *keyed*. From these we resynthesized new tokens having different *f*0 contours. Resynthesis was accomplished using ILS software on a VAX computer. Contours were flat with *f*0 at 96, 99, 102, 105, 108, and 111 Hz. (We used flat contours because they provided the most straightforward way to vary *f*0 and because Silverman, 1987, found little difference in judgments of flat and sloping contours in his research.) We next created a test order in which each token of *cod* was paired with each token of *keyed*. The pairings were made so that subjects heard all combinations of frequencies four times with the order *cod* and *keyed* counterbalanced. This gave 144 pairs of tokens (six *cod* frequencies × six *keyed* frequencies × two orders of *cod* and *keyed* × two tokens of each combination). There were 750 msec between words within a trial and 3.5 sec between trials. Seven seconds followed trials that corresponded to the end of a column on subjects' answer sheets.

Procedure. Answer sheets offered response alternatives *cod* and *keyed* for each trial. The subjects were instructed to decide which word in the pair they had heard on each trial was higher in pitch and to circle their answer on the sheets provided. We instructed the subjects to guess even if they did not detect a difference between the tokens.

Results and Discussion

We excluded data from 6 subjects whose response patterns were markedly nonmonotonic across the continuum.[1] Figure 2 shows the proportion of judgments that *keyed* is higher in pitch than *cod* averaged across the remaining 16 subjects and presented as a function of the *f*0 difference (/i/ − /a/) between the two words. We used probit analysis to fit ogives to the curves of individual subjects.

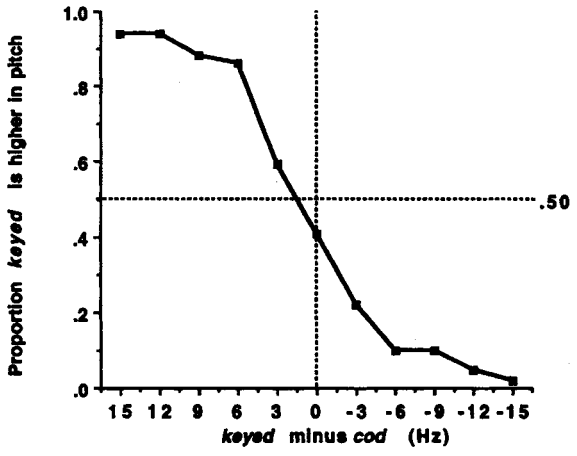

Figure 2. Proportion of judgments in which *keyed* exceeded *cod* in pitch as a function of the hertz difference between the vowels (data from Experiment 2).

Our dependent measure was the $f0$ difference between /i/ and /a/ at which, on the fitted ogive, subjects judged *keyed* as higher on 50% of opportunities. This value ranged from −1.22 to 3.42 across subjects and averaged 1.14 Hz. This is a significant departure from 0 [$t(15) = 2.87, p = .012$], and it signifies that, to sound equal in pitch to /a/, /i/ must exceed /a/ by 1.14 Hz on the average.[2]

Although our subjects showed significant evidence of parsing, as we predicted, they showed considerably less than expected—indeed, less than 1/10 (7%) the magnitude of the expected effect based on the average $f0$ difference produced by our talkers in Experiment 1. Of course, this may or may not signify that, outside the laboratory, subjects only parse 7% of the $f0$ perturbation from the intonation contour that talkers produce. We can probably reject the possibility that vowel-intrinsic $f0$ as produced in ordinary speech is considerably smaller in magnitude than that produced in the laboratory. Ladd and Silverman (1984) estimated a magnitude of $f0$ in fluent speech that is similar to the magnitude we found for the speech condition of Experiment 1. Factors that may lead us to underestimate the magnitude of parsing outside the laboratory include the nature and difficulty of the experimental task and the range of frequency differences we used.

Some evidence suggests that task difficulty may have reduced our estimate of parsing. When we correlate the value of the /a/–/i/ frequency difference at 50% responding for each subject with his/her standard error (representing the fit of the ogive to the data), we obtain a marginally significant correlation ($r = -.49, p = .055$). This negative correlation suggests that subjects who showed the largest positive frequency differences also tended to be those who responded most consistently. Our noisiest subjects produced the unexpected negative frequency differences. Indeed, of the 5 subjects who showed negative differences at 50% responding, the 3 with the largest negative differences also had the highest standard errors of the 16 subjects. Obviously, however, this cannot be the only factor, or even the most important one in determining the small amount of parsing that we saw in this experiment. The listener who parsed the most only ascribed 3.42 Hz of the $f0$ contour to intrinsic $f0$.

A second source of underestimation may be the range of frequency differences that we used. Our production experiment suggested that perfect parsers would parse 13–14 Hz from the $f0$ contour on /i/ vowels. The maximum frequency differences subjects received in Experiment 2 was 15 Hz. If they were perfect parsers, on the vast majority of trials they would judge *cod* to be higher in pitch than *keyed*. Possibly, listeners felt some constraint against always judging *cod* to be higher. Silverman (1987), who found considerably more parsing than we obtained, also used a much wider range of frequency differences than we did. Accordingly, in Experiment 3, we doubled the range of $f0$ differences among tokens of *cod* and *keyed*.

EXPERIMENT 3

The manipulation of frequency-difference range in Experiment 3 in comparison with that of Experiment 2 may have three possible outcomes. Our estimate of parsing may be unchanged, increase, or decrease. If Experiment 2 accurately estimated parsing, then increasing the range of frequency differences in Experiment 3 should not change our estimate of the magnitude of parsing. If, instead, the range of frequency differences that we used led us to underestimate parsing because it was too restricted, as we suggested above, then the increase in range should lead to an increase in our estimate of parsing and perhaps bring it close to the magnitude of intrinsic $f0$ differences between spoken /i/ and /a/ that we found in Experiment 1. The third logically possible outcome is a decrease in the estimate, but there is no reason that is obvious to us why this should occur.

Method

Subjects. The subjects were 18 undergraduate students who participated in the experiment for course credit. All were native speakers of English with normal hearing.

Materials. We resynthesized M1's spoken productions of *cod* and *keyed*. Contours were flat with $f0$ at 94, 100, 106, 112, 118, and 124 Hz. This created a maximum frequency difference of 30 Hz between the members of a word pair on each trial. This doubled the differences used in Experiment 2 and exceeded by approximately 17 Hz the intrinsic $f0$ difference in production that we observed in Experiment 1. The test order was identical to that used in Experiment 2, with the new values of $f0$ substituted for the ordinally corresponding ones of Experiment 2's test order.

Procedure. The procedure was identical to that of Experiment 2.

Results and Discussion

Judgments in this experiment were generally easier than those in Experiment 2 because the frequency differences between words in a pair were, on the average, twice those in Experiment 2. Data from no subject would have been excluded on the basis of the paired criteria of Experiment 2 (see note 1). However, 3 subjects did stand out

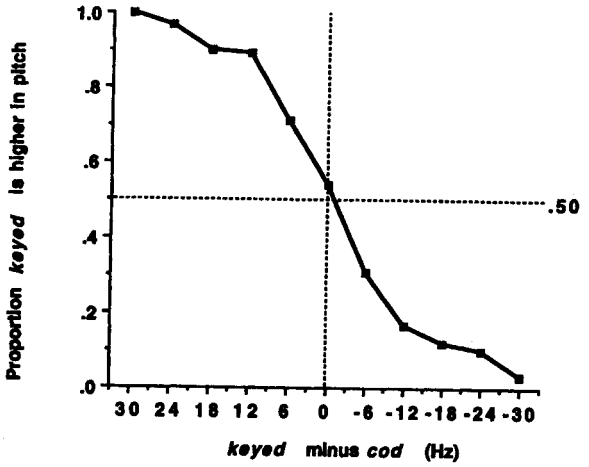

Figure 3. Proportion of judgments in which *keyed* exceeded *cod* in pitch as a function of the hertz difference between the vowels (data from Experiment 3).

markedly from the others in showing six changes in direction in the magnitude of their response proportions across the 11-item continuum. We excluded the data from these 3 listeners. Figure 3 presents the findings averaged across the remaining 15 participants. As in Experiment 2, we used probit analysis to fit ogives to the curves of individual subjects. In Experiment 2, the difference in fundamental frequency between /i/ and /a/ at which listeners did not distinguish the vowels in pitch averaged 1.14 Hz. In the present experiment, it averaged .32 Hz, a numerical reduction in the estimate of parsing. This value differs neither from 0—that is, no parsing [$t(14) = .54, p = .60$]—nor from 1.14 Hz [$t(29) = 1.15$, $p = .26$]. That is, parsing was statistically absent in the data of this experiment; however, it did not differ significantly in magnitude from the (significant) parsing that we found in Experiment 2. When we pooled the data across the experiments, our estimate of parsing remained significant [$t(30) = 2.09, p < .05$].

The correlation between intercept and standard error did not approach significance in this experiment; accordingly, we have no evidence, as we did in Experiment 2, that the most consistent responders parsed more from the $f0$ contour than did the less consistent responders.

The outcome of this experiment was surprising to us in showing a numerical reduction in the estimate of parsing under conditions that we expected would increase it or, at the very least, leave it unchanged. Statistically, of course, it was unchanged, and, for the present, we adopt that characterization of our findings, pending a second examination of the effects of manipulating the frequency-difference range, now on our estimate of parsing of sung vowels. Experiment 4 provides that examination.

EXPERIMENT 4

In the present experiment, we used resynthesis on the sung vowels of the vowel-shifting condition of Experiment 1 to examine parsing of intrinsic $f0$ from the perceived pitch of sung speech. Grieffenberg and Reinholt-Petersen (1982) reported that 1 listener (a staff member at an institute of musicology and so, in the investigators' judgment, "a highly competent listener") showed evidence of parsing. Our experiment tested for parsing among unselected listeners and compared its magnitude to that of the intrinsic $f0$ difference produced by subjects in Experiment 1. Accordingly, we expected parsing to average approximately 3 Hz (i.e., the $f0$ difference /i/ − /a/ at which listeners do not distinguish the vowels in pitch).

Method

Subjects. Thirty-six introductory psychology students participated in the experiment for course credit. All 36 students were native speakers of English with normal hearing. Of these, 17 were run in the narrow-range frequency-difference condition and 19 in the wide-range condition.

Materials. Recordings of M1 were used for this experiment, as for Experiment 2. We used single tokens of sung /a/ and /i/ from the vowel-shifting condition of Experiment 1. For this purpose, we chose approximately 625-msec ranges from one token of a trial in which the subject had shifted from sung /a/ to /i/. We excised these portions from the original production, giving us separate /a/ (624 msec) and /i/ vowels (627 msec).

These vowels were resynthesized to have flat $f0$ contours with $f0$ values of 96, 99, 102, 105, 108, and 111 Hz for the narrow-range condition and values of 94, 100, 106, 112, 118, and 124 Hz for the wide-range condition. These are the values we used in Experiments 2 and 3.

We made two audiotapes, one per frequency range, in which resynthesized sung vowels from the vowel-shifting condition of Experiment 1 were paired following the procedure of Experiments 2 and 3. Thus, in each test, there were 144 trials in which two tokens each of all pairings of /i/ and /a/ frequencies occurred and with the order of /i/ and /a/ counterbalanced. Timing within and between trials was as in Experiments 2 and 3.

Procedure. Subjects were assigned to one of two (frequency-difference range) groups. For each perception test, answer sheets provided the response options *ah* and *ee*. Subjects were instructed to decide which vowel in the pair they had heard on each trial was higher in pitch and to circle their answer on the sheets provided. As in the speech tests of Experiments 2 and 3, they were instructed to guess rather than to leave blanks.

Results

Vowel matching, narrow-frequency range. Data from 1 subject were eliminated using the criteria specified in note 1. The top panel of Figure 4 presents the data averaged over the remaining 16 subjects. As for Experiments 2 and 3, we fit an ogive to the data of each subject and used it to determine the frequency difference between /a/ and /i/ at which subjects judged the two vowels to have the same pitch. Averaged across subjects, this value was 3.35 hertz, a value very close to the predicted one. The 95% confidence limits surrounding this mean are 2.33 and 4.90 hertz; this interval includes the measured intrinsic $f0$ differences in the vowel-shifting condition of Experiment 1 of 3.0 Hz ($n = 13$) and 2.63 Hz ($n = 8$). Fifteen of the 16 subjects had positive hertz differences at the 50% point; accordingly, the difference between /a/ and /i/ ($f0$ of /i/ − $f0$ of /a/) was highly statistically significant [$t(15) = 4.30, p = .0006$]. In contrast to the re-

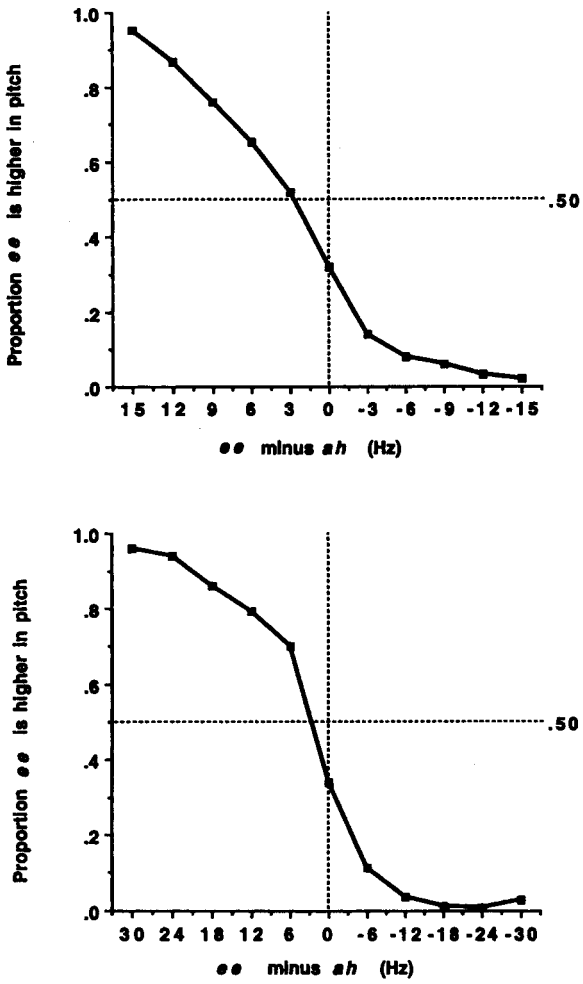

Figure 4. Proportion of judgments in which *ee* exceeded *ah* in pitch for vowels originally sung in isolation. The top and bottom panels, respectively, display data from the narrow- and wide-frequency difference conditions of Experiment 4.

sult in Experiment 2, the correlation between the hertz difference at the 50% responding point and the standard error reflecting the fit of subjects' data to the ogive did not approach significance. However, this is because, in this experiment, but not in Experiment 2, different subjects' magnitudes of parsing differed from the predicted amount both by showing less than the predicted amount of parsing and by showing more than the expected amount. A correlation between each subject's standard error and the absolute value of the departure of his/her parsing score from the predicted 3.0-Hz parsing was highly significant ($r = .78, p = .0001$). That is, subjects whose estimated magnitude of parsing was closest to 3 Hz also gave the best fits to the ogival function. As for individual data, the only subject to show a negative hertz difference also had the largest standard error. Further, the 3 subjects with identical, and the lowest, standard errors all had hertz differences close to the predicted value (2.25, 2.29, and 3.15). In short, here, in contrast to Experiments 2 and 3, on the average subjects parsed accurately. And more strongly than in Experiment 2, the best subjects, as estimated from standard errors, came closest to parsing accurately.

Vowel matching, wide-frequency range. Data from 1 subject were excluded on the basis of the criteria specified in note 1; no other subject's data met the more liberal exclusionary criteria of Experiment 3. The bottom panel of Figure 4 presents findings averaged across the 18 listeners. The frequency difference between /i/ and /a/ at the point of subjective pitch equality was 4.35 Hz, a value 1 Hz higher, but not significantly higher, than that found with the narrow range of frequency differences [$t(32) = .743, p = .46$]. Most of the difference between the intercepts in the two frequency range conditions can be ascribed to 1 subject with an intercept of nearly 18 Hz in the wide-range condition. As in the narrow-range condition, all but 1 subject had positive hertz differences at the 50% point; accordingly, the difference between /i/ and /a/ ($f0$ of /i/ $-$ $f0$ of /a/) was highly statistically significant [$t(17) = 4.09, p = .0008$]. The correlation between each subject's standard error and the absolute value of the departure of their parsing score from the predicted 3.0-Hz parsing was highly significant ($r = .74, p = .0002$). The listener with the largest departure from 3.0 Hz had the largest standard error; the listener with the smallest departure had the second smallest standard error.

Discussion

In contrast to findings of Experiments 2 and 3, the present experiment revealed a close match between the magnitude of parsing by listeners and the magnitude of intrinsic $f0$ differences measured in Experiment 1. Further, we found no effect of the range of frequency differences between /i/ and /a/ vowels across trials on the estimate of parsing.

As for the accuracy of parsing, for now, the most we can say is that, for the sung, isolated vowels of Experiment 4, listeners were accurate. For the spoken vowels in the word context of Experiments 2 and 3, they were not. We consider reasons for these different outcomes in the General Discussion. Experiment 5 provides a final estimate of parsing accuracy using vowels sung by M1 in the tone-matching condition of Experiment 1.

EXPERIMENT 5

Method

Subjects. The subjects were 14 students in an introductory psychology course who participated for course credit. They were native speakers of English with normal hearing.

Materials. For the tone test, we used the wider range of vowel fundamental frequencies of Experiments 3 and 4, and we synthesized a 640-msec sine wave tone at 109 Hz, the middle of the range of vowel frequencies. We selected tokens of vowels /i/ and /a/ from M1's tone condition. As in the previous experiment, we selected stretches in each vowel that were steady-state in $f0$. The /a/ vowel was 636 msec long and the /i/ vowel was 630 msec long. Vowels were resynthesized as in Experiment 4. In the test order presented

to listeners, the tone was paired with each of the resynthesized vowels, with the order of tone and vowel counterbalanced. There were 96 trials in all (six vowel frequencies × two vowels × two orders of tone and vowel × four tokens of each trial type). Timing within and between trials was as in Experiments 2–4.

Procedure. The subjects were asked to judge on each trial whether the tone or the vowel was higher in pitch. They responded by circling "tone" or "vowel" next to each trial number on their answer sheet. They were asked to circle exactly one of these choices on each trial, guessing if necessary.

Results and Discussion

Although data were somewhat messy across the board in this experiment, no subset of listeners stood out as especially deviant in their responding. Accordingly, data from all listeners were retained. Figure 5 shows the data averaged over the 14 subjects. The figure displays the frequency difference between the vowel and the tone (tone minus vowel) on the horizontal axis and the proportion of judgments for which the tone was higher in pitch than the vowel on the *y*-axis. The parameter in the figure is the identity of the vowel being judged relative to the tone. At every frequency difference, the /i/ curve lies above the /a/ curve, as predicted. That is, in all comparisons, the tone was more likely to be judged higher than /i/ than to be judged higher than /a/. In an analysis of variance with factors vowel and frequency difference, both main effects and the interaction were significant [vowel, $F(1,13) = 9.85$, $p = .008$; frequency difference, $F(5,65) = 22.54$, $p < .0001$; interaction, $F(5,65) = 4.04$, $p = .003$]. The interaction was significant because the vowel difference was smaller at frequency differences 15 and 9 Hz than elsewhere.

The data in the figure are not ogival. Accordingly, to estimate the amount of parsing, we fit regression lines rather than ogives to the curves. The fit of a line to the raw /i/ data was modest [$R^2 = .33$, $F(1,82) = 40.59$, $p < .0001$]. Setting the proportion of judgments for which the tone was higher to .5 and solving for the /i/–tone frequency difference at that point yielded a 5.53-Hz difference such that /i/ must be 5.53 Hz higher than the tone to be judged equal to it in pitch. The analogous analysis applied to the /a/ data gave a better fit to the line [$R^2 = .59$, $F(1,82) = 116.42$, $p < .0001$]. Solving for the /a/–tone frequency difference at subjectively equal pitch yielded a difference of −3.67 Hz such that /a/ matched the tone in pitch when it was lower in frequency than the tone by 3.67 Hz. If subtracting the outcome for /a/ from that for /i/ is the appropriate way to estimate parsing due to intrinsic *f*0, our estimate is 9.2 Hz, a value that is too high for the amount of intrinsic *f*0 that our singers produced in the vowels.

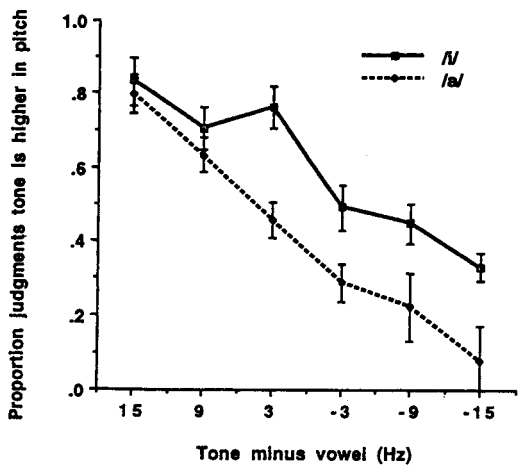

Figure 5. Proportion of judgments in which the tone was higher than *ee* or *ah* in Experiment 5. Standard error bars are shown.

We cannot look at individual accuracies in the way that we did in the speech- and vowel-matching experiments to determine whether the most consistent subjects were also the most realistic parsers. Data from 4 subjects in this experiment did not provide significant fits to a line. If we restrict our examination, however, to the 10 subjects for whom fits to lines yielded significant R^2s for both vowels, we do see lower estimates of parsing for these subjects than for the group of 14 subjects as a whole. Estimating parsing for each of these subjects in the way that we did for the group as a whole yielded estimates of intrinsic *f*0 parsing that averaged 4.26. This is much closer to the predicted value of about 2 Hz than the set of 14 subjects showed. Therefore, we do have evidence here, as in Experments 2 and 4, that the most systematic responders are the most accurate parsers.

GENERAL DISCUSSION

Our research was designed to address two general questions. First, does "parsing" of *f*0 occur in perception of speech and of sung speech when vowels differ in their intrinsic *f*0? Second, does the magnitude of *f*0 parsed from the perceived pitch of a vowel correspond to the magnitude of the intrinsic *f*0 difference? We can answer the first question in the affirmative. In five of five comparisons numerically and in four of five significantly, we found evidence that, matched in *f*0, an /i/ vowel sounds lower in pitch than an /a/ vowel whether it is spoken or sung.

We did not succeed in providing the affirmative answer we expected to the second question. Parsing was accurate for vowels that subjects had sung in isolation. In two comparisons that differed in the range of frequency differences between /a/ and /i/ pairs, parsing averaged 3.9 Hz, closely matching the measured intrinsic *f*0 difference of 3 Hz. However, we did not find a close match between measured intrinsic *f*0 differences and parsing among spoken words. There, the measured difference was 13–14 Hz, but the magnitude of parsing was 1.14 Hz in one experiment and a nonsignificant .32 Hz in another. Finally, vowels sung to match a tone differed by 2 Hz on the average; however, our listeners judged /i/s and /a/s as matching the tone when they differed from each other by 9 Hz. Accordingly, across experiments, we have listeners parsing too little, parsing accurately, and parsing excessively.

In four of five comparisons (except the comparison made in Experiment 3), we have evidence that the most consistent responders among our listeners were the most accurate parsers. In addition, particularly among our most consistent responding listeners, parsing was more accurate for sung than for spoken vowels. We next consider why our findings might have differed between the spoken and sung conditions.

One possibility is that parsing really is more accurate for sung than for spoken vowels. The requirement that sung speech be perceived as matching tones from any instrumental accompaniment may foster more accurate parsing than in speech perception. However, although this may provide part of the answer, we doubt that it provides all of it. First, it is not obvious why listeners would parse $f0$ from the intonation contour of a spoken utterance at all if their parsing amounted to only 7% of the "distortion" to the intonation contour provided by intrinsic $f0$ (as we found when we compared intrinsic $f0$ magnitudes in Experiment 1 to parsing in Experiment 2). This leads us to guess that our research methods underestimate spoken-word $f0$ parsing. Compatible with this interpretation, Silverman (1987) reported a considerably larger magnitude of parsing than we found here.

There are some procedural differences between Silverman's (1987) study and our own. Most notably, he presented his /i/ and /a/ vowels, not only in real words as we did here, but also in sentence context (e.g., "They only feast before fasting," and "They only fast before feasting" produced by a speaker of British English). This may render more salient and important recovery of the intontation contour of the utterance and may foster parsing. Somewhat compatibly, in a different study of parsing of $f0$ perturbations on a vowel due to the voicelessness of a preceding consonant (Pardo & Fowler, in press), we did find a tendency for parsing to be larger when words are presented in a sentence context than when words occur in isolated pairs, as in Experiments 2–5. However, the difference was not large enough to explain the discrepancy between the magnitude of intrinsic $f0$ measured in Experiment 1 (and elsewhere; see Silverman, 1987, for a review) and the parsing we saw in Experiments 2 and 3. We hope in future investigations to develop a better understanding of procedural factors that affect our estimates of parsing of $f0$. We do speculate that parsing is more accurate than our studies reveal, on grounds that inaccurate parsing would appear to be no more useful than no parsing at all, and except Experiment 3, we have found reliable evidence that parsing does occur.

The occurrence of parsing provides a window on the nature of speech perception. It is quite revealing that listeners do not treat a unitary physical dimension of the acoustic signal such as $f0$ as if it were *informationally* unitary. From the perspective of one type of theory of speech perception, $f0$ tends not to be informationally unitary. This is a type of theory, including the motor theory of speech perception (Liberman & Mattingly, 1985) and the direct-realist theory (Best, 1994; Fowler, 1986), in which the acoustic signal is supposed to provide information for its causal sources, namely phonetic gestures of the vocal tract. When different phonetic gestures have converging effects on common acoustic dimensions, here on fundamental frequency, listeners who use the acoustic signal as information for gestures should parse those dimensions as if into their distinct informational components. In our research, the vowel gesture and the gesture producing the fundamental frequency contour of spoken words or of sung vowels had converging effects on fundamental frequency. Accordingly, for a perceiver of gestures, extraction of acoustic information for /i/ should include detection of /i/ information in $f0$ on the vowel. The perceived pitch of the vowel should then correspond to the residual $f0$ not ascribed to the vowel gesture.

In the theory of direct perception, in particular, listeners are predicted to parse accurately, at least in the case of natural speech signals, in which gestural sources are presumed to be specified acoustically. If our resynthesized signals also provided specifying information, as we intended them to, then the inaccuracies of parsing are not supportive of the theory. In defense of the theory, however, we point out that the occurrence of parsing has not been predicted at all by any other theorists, and our findings very clearly show that parsing of $f0$ occurs. Further, it is difficult to imagine that other theories, however they might be developed, would predict the occurrence of parsing, but inaccurate parsing. Parsing is useful only if it permits perception of the phonetic properties of an utterance that a speaker intended to convey, in the case of phonetic properties signaled in part by $f0$, the speaker's intended intonation contour, intended vowel height, intended stress, and intended consonant voicing. Inaccurate parsing in which listeners parse sometimes too much and sometimes too little $f0$ should be no more useful than no parsing at all, but parsing does occur.

REFERENCES

BEST, C. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. Goodman & H. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167-224). Cambridge, MA: MIT Press.

DIEHL, R. (1991).The role of phonetics within the study of language. *Phonetica*, **48**, 120-134.

DHYR, N. (1990). The activity of the cricothyroid muscle and the intrinsic fundamental frequency in Danish vowels. *Phonetica*, **47**, 141-154.

FISCHER-JØRGENSEN, E. (1990). Intrinsic $f0$ in tense and lax vowels with special reference to German. *Phonetica*, **47**, 99-140.

FOWLER, C. A. (1981). Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech & Hearing Research*, **46**, 127-139.

FOWLER, C. A. (1984). Segmentation of coarticulated speech in perception. *Perception & Psychophysics*, **36**, 359-368.

FOWLER, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, **14**, 3-28.

FOWLER, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, **99**, 1730-1741.

FOWLER, C. A., & SMITH, M. (1986). Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In J. Perkell & D. Klatt (Eds.), *Invariance and variability of speech processes* (pp. 123-136). Hillsdale, NJ: Erlbaum.

GRIEFFENBERG, M., & REINHOLT-PETERSEN, N. (1982). *The effect of high and low vowels on the fundamental frequency in singing: Some*

preliminary observations (Annual Report). Copenhagen: University of Copenhagen, Institute of Phonetics.

HOMBERT, J. M. (1978). Consonant types, vowel quality and tone. In V. Fromkin (Ed.), *Tone: A linguistic survey* (pp. 77-112). New York: Academic Press.

HONDA, K. (1981). Relationship between pitch control and vowel articulation. In D. Bless & J. Abbs (Eds.), *Vocal-fold physiology* (pp. 286-297). San Diego: College-Hill.

HONDA, K., & FUJIMURA, O. (1991). Intrinsic vowel *F*0 and phrase-final *F*0 lowering: Phonological versus biological explanations. In J. Gauffin & B. Hammarberg (Eds.), *Vocal fold physiology: Acoustic perceptual and physiological aspects of voice mechanisms* (pp. 149-157). San Diego: Singular.

KLUENDER, K. (1994). Speech perception as a tractable problem in cognitive science. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 173-217). San Diego: Academic Press.

LADD, D. R., & SILVERMAN, K. E. A. (1984). Vowel intrinsic pitch in connected speech. *Phonetica*, **41**, 31-40.

LIBERMAN, A., & MATTINGLY, I. (1985). The motor theory revised. *Cognition*, **21**, 1-36.

MARTIN, J., & BUNNELL, H. T. (1981). Perception of anticipatory coarticulation effects. *Journal of the Acoustical Society of America*, **69**, 559-567.

OHALA, J., & EUKEL, B. (1976). *Explaining the intrinsic pitch of vowels* (Working paper). Berkeley: University of California, Berkeley Phonology Laboratory.

OHALA, J., & EUKEL, B. (1987). Explaining intrinsic pitch of vowels. In R. Channon & L. Shockey (Eds.), *In honor of Ilse Lehiste* (pp. 207-215). Dordrecht: Foris.

PARDO, J., & FOWLER, C. A. (in press). Perceiving the causes of coarticulatory acoustic variation. *Perception & Psychophysics*.

REINHOLT-PETERSON, N. (1986). Perceptual compensation for segmentally-conditioned fundamental-frequency perturbations. *Phonetica*, **43**, 31-42.

SAPIR, S. (1989). The intrinsic pitch of vowels: Theoretical, physiological and clinical observations. *Journal of Voice*, **3**, 44-51.

SILVERMAN, K. (1987). *The structure and processing of fundamental frequency contours*. Unpublished doctoral dissertation, Cambridge University.

TERNSTRÖM, S., SUNDBERG, J., & COLLDÉN, A. (1988). Articulatory F_0 perturbations and auditory feedback. *Journal of Speech & Hearing Research*, **31**, 187-192.

VILKMAN, E., AALTONEN, O., LAINE, U., & RAIMO, I. (1991). Intrinsic pitch of vowels—A complicated problem with an obvious solution? In J. Gauffin & B. Hammarberg (Eds.), *Vocal fold physiology: Acoustic perceptual and physiological aspects of voice mechanisms* (pp. 159-166). San Diego: Singular.

VILKMAN, E., AALTONEN, O., RAIMO, I., ARAJÄRVI, P., & OKSANEN, H. (1989). Articulatory hyoid-laryngeal changes vs. cricothyroid activity in the control of *F*0 of vowels. *Journal of Phonetics*, **17**, 193-203.

WHALEN, D. H. (1984). Subcategorical mismatches slow phonetic judgments. *Perception & Psychophysics*, **35**, 49-64.

WHALEN, D. H., & LEVITT, A. (1995). The universality of intrinsic *F*0 of vowels. *Journal of Phonetics*, **23**, 349-366.

WHALEN, D. H., LEVITT, A., HSAIO, P., & SMORODINSKY, I. (1995). Intrinsic *F*0 of vowels in the babbling of 6-, 9-, and 12-month old French- and English-learning infants. *Journal of the Acoustical Society of America*, **97**, 2533-2539.

NOTES

1. We excluded subjects who met the conjunction of two criteria. First, at the ends of the continuum, response proportions were farther from 0 than .25 at one end or farther from 1.00 than .75 at the other; second, more than four changes in direction in the magnitude of their response proportions occurred across the 11-item continuum.

2. In Figures 2–4, the 50% intercept apparent in the figures may not exactly match the average intercept we report in the text. In the text, we report values obtained by fitting ogives to the data of individual subjects and then averaging the intercepts across subjects. These are the values on which statistical tests were performed. Figures present data points averaged over subjects.

(Manuscript received April 18, 1996;
revision accepted for publication August 13, 1996.)