Attention, Perception, & Psychophysics
2009, 71 (5), 1150-1166
doi:10.3758/APP.71.5.1150

The role of f_0 and formant frequencies in distinguishing the voices of men and women

James M. Hillenbrand and Michael J. Clark
Western Michigan University, Kalamazoo, Michigan

The purpose of the present study was to determine the contributions of fundamental frequency (f_0) and formants in cuing the distinction between men's and women's voices. A source-filter synthesizer was used to create four versions of 25 sentences spoken by men: (1) unmodified synthesis, (2) f_0 only shifted up toward values typical of women, (3) formants only shifted up toward values typical of women, and (4) both f_0 and formants shifted up. Identical methods were used to generate four corresponding versions of 25 sentences spoken by women, but with downward shifts. Listening tests showed that (1) shifting both f_0 and formants was *usually* effective (~82%) in changing the perceived sex of the utterance, and (2) shifting either f_0 or formants alone was usually *ineffective* in changing the perceived sex. Both f_0 and formants are apparently needed to specify speaker sex, though even together these cues are not entirely effective. Results also suggested that f_0 is somewhat more important than formants. A second experiment used the same methods, but isolated /hVd/ syllables were used as test signals. Results were broadly similar, with the important exception that, on average, the syllables were more likely to shift perceived talker sex with shifts in f_0 and/or formants.

It is well-known that speech signals convey a great deal of information in addition to the linguistic features that have understandably attracted the largest share of attention in the speech perception literature. This extralinguistic information includes features that allow listeners to distinguish men's voices from women's voices. The most obvious and heavily studied candidates for conveying speaker sex information are differences in fundamental frequency (f_0) and formant frequencies. Typical fundamental frequencies are slightly less than an octave lower for men than for women and are the result of the longer and heavier vocal folds that are usually observed in men (see Titze, 1989, for a review). Values differ somewhat across studies, but the averages reported by Hillenbrand, Getty, Clark, and Wheeler (1995) for 1,116 /hVd/ utterances spoken by 45 men and 48 women are typical, with a mean of 131 Hz for the men and 220 Hz for the women, a difference of 0.75 octave. Nearly identical figures of 132 and 224 Hz were reported by Peterson and Barney (1952), based on 1,220 /hVd/ utterances from 33 men and 28 women. Differences in formant frequencies are the result of the somewhat longer vocal tracts typical for men. Scale factors (women/men) derived from the Hillenbrand et al. (1995) data, averaged across all vowels, are 1.18, 1.17, and 1.14 for F_1, F_2, and F_3, respectively. Comparable figures from Peterson and Barney are 1.16, 1.19, and 1.16.

A priori, there would seem to be reasons for speculating that f_0 might serve as a more compelling cue to speaker sex than formant frequencies. The most obvious consideration is that the male–female difference in f_0 is proportionally much larger than the typical differences in formant frequencies. It is also the case that f_0 can be evaluated largely independently of the phonetic identity of the speech sound being spoken, whereas making use of formant frequency information depends heavily on knowing the identity of the speech sound. For example, an F_1 value of 750 Hz might suggest either /ɑ/ (or perhaps /ʌ/) spoken by a man or /ɔ/ spoken by a woman. This gives rise to a circularity that is seldom discussed: Pattern recognition studies have shown that vowels can be categorized with considerably greater accuracy with the inclusion of f_0 as a normalizing parameter (e.g., Disner, 1980; Hillenbrand & Gayvert, 1993; Hillenbrand et al., 1995; J. D. Miller, 1989; Nearey, 1978; Nearey, Hogan, & Rozsypal, 1979; Syrdal & Gopal, 1986). There is also ample evidence that f_0 plays a significant role in listener judgments of vowel identity (e.g., Ainsworth, 1975; Fujisaki & Kawashima, 1968; R. L. Miller, 1953; Nearey, 1989; Potter & Steinberg, 1950; Slawson, 1968). However, the reverse is also true: As is noted below, speech samples spoken by men and women can be statistically separated on the basis of formant frequencies with far greater accuracy if the identity of the speech sound is known. Furthermore, as is also discussed below, there is some evidence suggesting that judgments of speaker sex may depend on the accuracy with which listeners judge vowel identity, and that judgments of vowel identity may depend on the accuracy with which listeners judge speaker sex (Eklund & Traunmüller, 1997).

J. M. Hillenbrand, james.hillenbrand@wmich.edu

Although f_0 and formants have received the most attention, these two features do not exhaust the possibilities. Having noted that synthetic speech modeled after adult male talkers did not shift to entirely convincing adult female voices after appropriate adjustments to f_0 and formant frequencies, Klatt and Klatt (1990) reported the results of acoustic measurements suggesting that, on average, women's voices tended to be slightly breathier than those of men. Acoustic and perceptual tests indicated that the primary correlates of increased breathiness in women's voices were reductions in signal periodicity, increased F_1 bandwidths, and decreased F_1 amplitudes, suggesting higher levels of aspiration noise. They also reported higher first harmonic amplitudes for women, suggesting more rounded glottal source waveforms (see also Childers & Wu, 1991). Some reports have suggested that prosodic variation is wider in women than in men (Fitch & Holbrook, 1970; Graddol & Swann, 1983; Hudson & Holbrook, 1981; Stoicheff, 1981), although there is some disagreement on this point (cf. Henton, 1989).

Pattern Classification

Bachorowski and Owren (1999) reported linear discriminant analysis results using acoustic measures made from the vowel /ɛ/ in the word *test*. Classification parameters consisted of f_0, a vocal tract length measure estimated from the frequencies of F_1–F_3, formant amplitudes, and periodicity (jitter and shimmer). f_0 and formant frequencies alone classified men and women quite accurately, with a relatively small advantage for f_0 (95.9%) versus formants (92.4%). f_0 and formants together separated men and women almost perfectly (98.8%). Formant amplitudes and periodicity contributed little to classification accuracy. It should be noted that the use of formant frequencies in separating men from women was greatly simplified by restricting formant measurements to a single vowel, making absolute formant frequencies considerably more informative than they would be in the more realistic case of varying phonetic identity. Additional pattern classification results on male–female differences are reported in the present study.

Listening Studies

A number of studies have demonstrated that listeners can accurately identify speaker sex from speech signals that are not phonated, evidence that is typically interpreted as implicating a role for male–female differences in formant frequencies. Schwartz and Rine (1968), for example, reported near-perfect judgments of talker sex from whispered vowels recorded from a small sample of men and women, and, using grade school-aged children as listeners, Bennett and Montero-Diaz (1982) reported reasonably accurate judgments of speaker sex for whispered vowels. Speaker sex can also be judged from isolated voiceless fricatives (Ingemann, 1968) and from sine wave replicas of short sentences (Fellowes, Remez, & Rubin, 1997). Sine wave replicas, being aperiodic, do not have pitch in the traditional sense.[1] The speaker-sex percept for these utterances is therefore almost certainly related mainly (although perhaps not exclusively; see below) to formant frequency differences. Finally, Coleman (1971) showed that women were still overwhelmingly heard as female even when they produced speech using an electrolarynx set at 85 Hz, well toward the low end of the adult male f_0 range.

There is also a good deal of evidence confirming a strong role for f_0 in the identification of speaker sex. For example, Bennett and Montero-Diaz (1982) and Lass, Hughes, Bowyer, Waters, and Bourne (1976) reported significantly better identification of speaker sex from phonated than from whispered vowels, suggesting that f_0 makes an important contribution. Coleman (1976) asked listeners to provide a graded measure of maleness or femaleness from time-reversed (i.e., played backward) connected speech samples originally recorded from 20 men and 20 women. Coleman (1976) reported that gender ratings were more strongly correlated with f_0 than with average formant frequencies. Lass et al. also reported very high identification rates for phonated vowels that were low-pass filtered in an attempt to eliminate formant cues. Noting that sex identification was more accurate for the presumably "formantless" low-pass filtered vowels (91%) than for "pitchless" whispered vowels (75%) from the same talkers, Lass et al. concluded that f_0 was a more potent cue to speaker sex. It is not at all clear, however, whether the low-pass filter used in that study was sharp enough to entirely eliminate formant information.

A second experiment by Coleman (1976) produced equivocal results. From his set of 20 men and 20 women, Coleman (1976) selected the 5 men with the lowest formant frequencies and the 5 women with the highest formant frequencies. These 10 talkers then produced one set of utterances with an electrolarynx generating a simulated glottal tone at 120 Hz and a second set with the electrolarynx generating a 240-Hz tone. For reasons that are not clear, listeners heard the majority of the key mismatched utterances (i.e., low f_0 combined with female formants and high f_0 combined with male formants) as men.

A few studies have used synthesis methods to assess the relative importance of f_0 and formants in cuing perceived speaker sex. A small-scale study by Whiteside (1998), for example, used a formant synthesizer to create four versions of 10 vowels: (1) formants and f_0 set to values typical of men (based on averages calculated from just three men), (2) formants and f_0 set to values typical of women (again, based on just three talkers), (3) f_0 set to values typical of men but with formants typical of women, and (4) f_0 set to values typical of women but with formants typical of men. Three listeners were asked to identify the test signals as male or female. Whiteside reported that, in the two key conflicting-cue conditions, f_0 was much more likely to be the dominant cue (i.e., signals with male f_0 and female formants were usually heard as men, and signals with female f_0 and male formants were usually heard as women).

Smith and Patterson (2005) recorded five sustained vowels (/i/, /e/, /ɑ/, /o/, /u/) from a single adult male, then used a source-filter synthesizer (Kawahara, Masuda-Kasuse, & de Cheveigné, 1999) to shift the f_0 to an exceedingly wide range of values between 61 and 523 Hz and to shift the vocal tract filters (and, consequently, the formants) to simulate an equivalently wide range of vocal tract lengths. Listeners were asked to provide judgments of speaker size and sex/age (i.e., *man*, *woman*, *boy*, *girl*) for each stimulus. Smith and Patterson concluded that listener judgments of both speaker size and sex/age were affected about equally by f_0 and formants. The distribution of sex/age judgments, however, corresponded only very approximately with values of f_0 and formant frequencies measured from natural utterances spoken by men, women, boys, and girls. Smith and Patterson also reported very uneven use of the four response categories, with "man" ($p = .36$) and "boy" ($p = .36$) responses occurring far more often than "woman" ($p = .11$) and "girl" ($p = .17$) responses. The small size of the woman category is especially striking and might be due to some combination of (1) the relatively large overlap between the f_0 and formant frequencies of women and children (although it is not clear why the boy and girl categories should spread to the woman category rather than vice versa) and (2) residual cues to maleness in the original recordings, made from an adult male talker.

A follow-up study by Smith, Walters, and Patterson (2007) used the same basic approach, but unlike Smith and Patterson (2005), which used one adult male as the basis for all rescaled signals, original recordings were made from a man, a woman, a boy, and a girl. As with the earlier study, it was generally the case that stimuli with lower values of f_0 and formant frequencies were heard as men, those with higher values as boys or girls, and those with intermediate values as women. Also consistent with the earlier study, there was only a rough correspondence between the shapes of the four response categories and f_0 and formant frequencies values measured from the speech of men, women, boys, and girls. The detailed findings across the four original talkers, however, were anything but simple. For example, the man and woman response categories were about the same for stimuli modeled after the man and the woman, but, for stimuli modeled after the child talkers, much lower values of f_0 and formant frequencies were needed to elicit "man" responses. Furthermore, although a clear *woman* category emerged from the signals modeled after the adult talkers, with signals modeled after the child talkers, *woman* disappeared as a dominant response category (with *dominant* being defined as a region enclosing response probabilities of at least .5).

A related study by Assmann, Nearey, and Dembling (2006) used the Kawahara et al. (1999) source-filter synthesizer to rescale the fundamental frequencies and spectrum envelopes of two sentences spoken by two men and two women. Fundamental frequencies were scaled to 10 values from 60 to 450 Hz, and vocal tract filters (and therefore formants) were scaled to 10 values covering a wide range corresponding to very short and very long vocal tracts. Listeners rated the 800 signals on a continuous scale ranging from *clearly masculine* to *clearly feminine*. As expected, signals with low fundamental frequencies and low formants were heard as decisively masculine, those with high values on both dimensions were heard as decisively feminine, and mismatched signals received intermediate ratings. Like Smith and Patterson (2005), Assmann et al. (2006) concluded that f_0 and formants contributed about equally to the perception of speaker sex. The results, however, were asymmetric: Ratings of masculinity for signals with downward frequency shifts were more pronounced than ratings of femininity for signals with equivalent upward shifts. Assmann et al. (2006) also found that, *even with equivalent f_0 and formant values*, signals synthesized from sentences originally spoken by men were more likely to be heard as masculine than were signals originally recorded from women. Conversely, signals synthesized from sentences originally spoken by women were more likely to be heard as feminine. As Assmann et al. (2006) noted, these results demonstrate that listeners are sensitive to cues to talker sex other than the scaling of f_0 and formant frequencies; in other words, the frequency-shifted utterances contained what Assmann et al. (2006) called "residual indicators of voice gender" (p. 892). It should be noted that this finding differs from that of Smith et al. (2007), who reported very similar distributions of "man" and "woman" responses to stimuli modeled after the recordings of a man and a woman. This discrepancy might indicate that residual cues to speaker sex are more readily conveyed by connected speech than by isolated vowels. This possibility is explored in the present study, which examines the effects of f_0 and formant shifting on both sentences and isolated syllables.

One final point on the perception of speaker sex concerns Owren, Berkowitz, and Bachorowski's (2007) hypothesis that listeners should be expected to show a perceptual advantage for male speech in tasks that involve the perception of speaker sex. Owren et al. argue the following:

> because sexual selection leads males to diverge from the "default" female form [i.e., physiological changes in speech structures occurring at puberty], adult male voices can be considered "marked" by the sexually selected features of lowered f_0 and formant frequencies. It therefore follows that listeners should hear talker sex somewhat more easily in male than in female voiced sounds. Specifically, the presence of critical features of "maleness" [low f_0, low formants] virtually guarantees that the talker is an adult male. However, their absence does not unequivocally imply that the talker is an adult female. (p. 930)

The authors point to higher recognition rates for men than for women in the findings of several studies (e.g., Bennett & Montero-Diaz, 1982; Coleman, 1971, 1976; Lass et al., 1976). The electrolarynx findings of Coleman (1976), showing that mismatched signals (high formants, low pitch; low formants, high pitch) were heard predominately as men, could also be cited here, along with the findings of Assmann et al. (2006) showing that ratings of masculinity for signals with downward frequency shifts were more pronounced than ratings of femininity for signals with equivalent upward shifts. In three experiments testing this "male advantage" hypothesis, Owren et al. (2007) reported that listeners identified brief, excised vowels spoken by men more accurately, and with shorter response latencies, than they did those of women. Finally, the authors reported the results of a male–female discrimination experiment using a signal detection paradigm, which they interpreted as ruling out response bias as an explanation for the male advantage that was observed in the three labeling experiments. Although the present study was not designed with this issue in mind, the findings reported here will have some limited bearing on this question.

In summary, there is clear evidence from both pattern classification studies and listening experiments confirming a significant role for both f_0 and formant frequencies in controlling the speaker sex percept, but no consensus has emerged on the relative importance of the two cues, either in relation to one another or in relation to other cues to speaker sex. The present study, which focused exclusively on distinguishing the voices of men and women, was designed to explore this issue further using a frequency-shifting approach similar to that used in Assmann et al. (2006) and Smith and Patterson (2005; Smith et al., 2007). The focus of these earlier studies on exploring a broad sampling of the f_0 and formant space, although entirely appropriate, made it necessary, for practical reasons, to base the synthesis on a small number of talkers (a single adult male in the case of Smith & Patterson, 2005; one example each of a man, a woman, a boy, and a girl in Smith et al., 2007; two men and two women in Assmann et al., 2006). Variability across talkers in a wide range of acoustic properties—including but not limited to f_0 and formants (e.g., Hillenbrand et al., 1995; Peterson & Barney, 1952)—is one of the most extensively documented facts in the acoustic phonetics literature. In the present study, the effects of frequency shifting on the speaker-sex percept were tested using many more talkers, but with a more restricted set of frequency shifts. A secondary goal was to compare the contributions of f_0 and formants to the perception of speaker sex using sentences and isolated syllables. Specifically, listeners were asked to identify speaker sex (and to provide confidence ratings) for utterances synthesized from 25 men and 25 women under several conditions. Recordings from the men were synthesized under four conditions: (1) f_0 and formants unmodified from their measured values, (2) f_0 and formants shifted up toward values typical of women, (3) f_0 alone shifted up toward values typical of women, and (4) formants alone shifted up toward values typical of women. Similarly, recordings from 25 women were used to synthesize signals under four corresponding conditions, but with frequencies shifted down toward values typical of men. Experiment 1 used sentences and Experiment 2 tested isolated /hVd/ syllables.

Discriminant Analysis

Method

As a preliminary to the synthesis experiments, pattern recognition tests were carried out in an effort to determine how well men's and women's speech can be distinguished on the basis of various combinations of f_0 and formant frequencies. A quadratic discriminant analysis technique was used for classification. Measurements of f_0 and F_1–F_3 were taken from 1,116 /hVd/ utterances consisting of 12 vowels (/i, ɪ, e, ɛ, æ, ɑ, ɔ, o, ʊ, u, ʌ, ɝ/) spoken by 45 men and 48 women (Hillenbrand et al., 1995, hereafter H95). A split-half method was used in which each category is trained on a random half of the utterances and tested on the remaining half.

The tests described here are similar to those reported by Bachorowski and Owren (1999), with the notable exception that measurements were made from 12 vowel categories rather than the single vowel type (/ɛ/) used in the earlier study. With variation across vowel category, the usefulness of absolute formant frequencies in differentiating the voices of men and women is reduced considerably since formant values are more strongly conditioned by vowel identity than by speaker sex.[2] The present tests were carried out with absolute frequencies and with normalized frequencies. The purpose of the normalization scheme was to represent frequency values not in absolute terms but in relation to values that are typical of the specific vowel that is spoken. In other words, the goal was to find a representation that indicates whether the formant frequencies are high or low not in absolute terms, but in relation to the formant frequencies that are typical of the speech sound that is being uttered, producing a measure reflecting primarily variations in vocal tract length and factoring out vowel identity as much as possible. We used a simple scheme in which frequency values were converted to standard scores using means and standard deviations (*SD*s) computed within vowel category but across men and women talkers. For example, a normalized F_1 value of 1.0 for a particular token simply indicates that the frequency value is 1 *SD* above the grand mean for F_1 *for the vowel that was spoken*. Figure 1 shows the distribution of absolute and normalized values of F_1–F_3 for men and women in the H95 database. The simple point made by panel A of this figure is that F_1 and F_2 are so strongly correlated with vowel identity that men and women are not easily distinguishable on the basis of their absolute frequencies. F_3, on the other hand, being on average much more loosely correlated with vowel identity, can be used to distinguish men from women reasonably well, although there is a fair amount of overlap. Panel B shows that men and women are reasonably distinct when normalized values are used. Figure 2 shows the distribution of absolute and normalized f_0. Unlike formant frequency, speaker sex is distinct in both representations.

Results

Discriminant analysis results are summarized in Table 1, which shows percent correct classification accuracy averaged over 30 random split-half tests. Results are shown for absolute and normalized frequencies using (1) f_0 alone, (2) various combinations of formants alone, and (3) f_0 and formants combined in various ways. It can be seen that f_0 alone does a very good job of separating men from women. The normalization of f_0 has little effect, which is unsurprising given that f_0 is heavily conditioned by speaker sex and, by comparison, only lightly conditioned by vowel identity. By contrast, normalization makes a great deal of difference when formants are used to distinguish speaker sex. Normalized formants distinguish speaker sex with reasonable accuracy, particularly when all three formants are used, whereas absolute formant frequencies perform rather poorly. Using all three normalized formants works best, and it does not seem

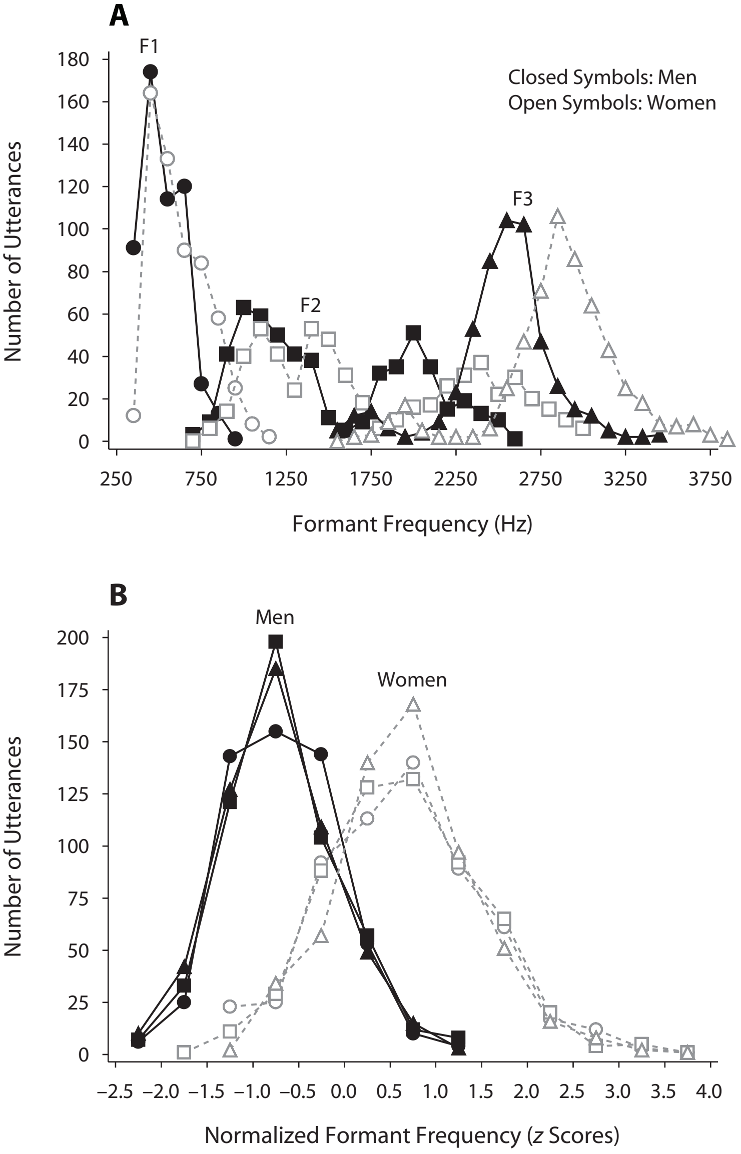

Figure 1. Frequency-of-occurrence histograms for absolute (A) and normalized (B) values of F_1 (circles), F_2 (squares), and F_3 (triangles) measured from 12 vowels spoken by 45 men (closed symbols) and 48 women (open symbols). Note that speaker sex is poorly distinguished on the basis of absolute formant frequencies, but reasonably well distinguished on the basis of normalized values.

to matter whether the individual formants are treated as separate parameters or whether they are averaged to form a single parameter. Normalized formants, however, do not separate men and women quite as well as f_0 does. The combination of f_0 and normalized F_1–F_3 separates men and women with considerable accuracy.

Conclusions drawn from the pattern recognition tests include the following: (1) f_0 alone distinguishes speaker sex with a high degree of accuracy (~96%). (2) The three lowest formant frequencies together distinguish speaker sex accurately (~92%) but only when those frequencies are normalized; that is, what is needed is a representation that indicates whether the formant frequencies are high or low in relation to the formant frequencies that are typical of the speech sound that is being uttered. (3) f_0 alone distinguishes speaker sex somewhat more accurately than do formants alone. (4) f_0 and formants in combination distinguish speaker sex more accurately than does either feature alone.

Statistical pattern recognition methods can be used to measure the degree of correspondence between perceptual categories and specific sets of physical features. As a consequence, pattern recognizers are useful in suggesting logically possible perceptual mechanisms, but, of course, a pattern recognizer is not a listener. The purpose of the experiments described below was to measure the relative contributions of f_0 and formants using synthetic speech in which these two features were manipulated separately and in combination.

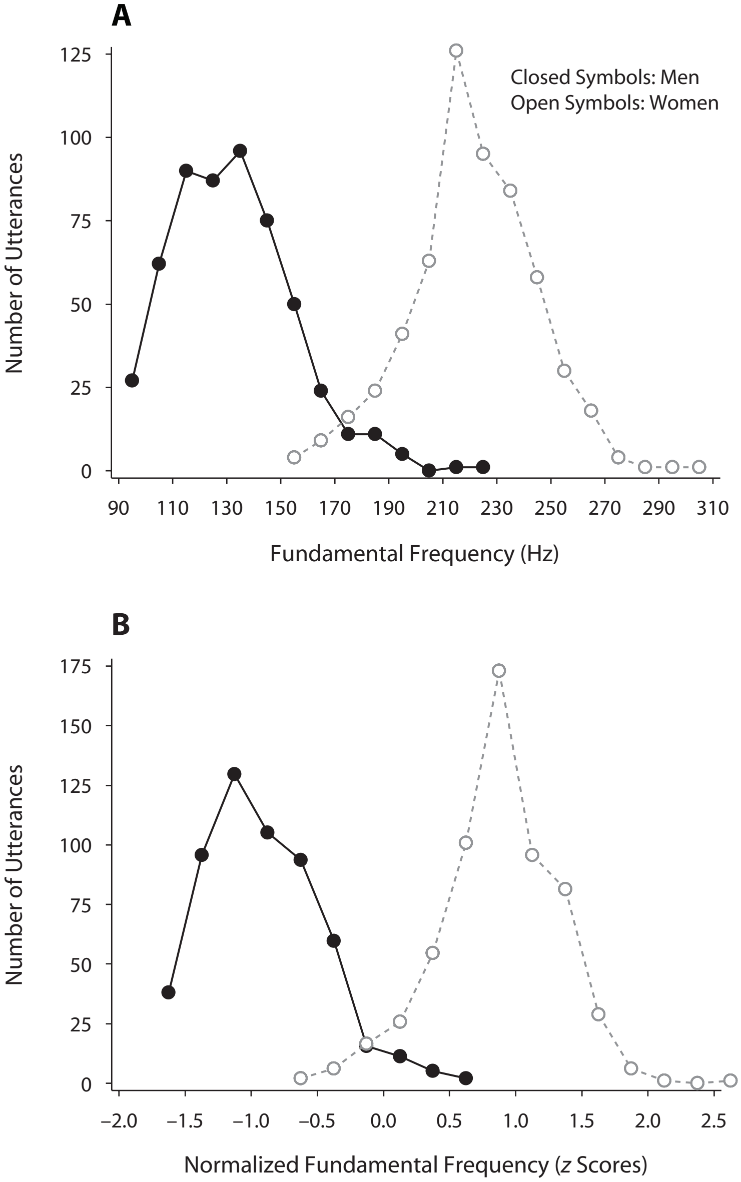

Figure 2. Frequency-of-occurrence histograms for absolute (A) and normalized (B) values of fundamental frequency (f_0) measured from 12 vowels spoken by 45 men (closed symbols) and 48 women (open symbols). Note that, unlike formant frequency, speaker sex is easily distinguished on the basis of either absolute or normalized values.

EXPERIMENT 1

Method

Sentences spoken by 25 men and 25 women were drawn at random from the phonetically diverse subset of the TIMIT (Texas Instruments–Massachusetts Institute of Technology) continuous speech database (Zue, Seneff, & Glass, 1990), unmodified from their original 16-kHz sample rate. A source-filter synthesizer (described below) was used to generate four versions of each sentence. For the men, the four versions were as follows: (1) MUS (men, unmodified synthesis): synthetic version of the original sentence with f_0 and spectrum envelope (and therefore formants) unmodified from their measured values. (2) MPE (men, pitch and envelope shifted): synthetic version with both f_0 and envelope shifted up toward values typical of women. (3) MPO (men, pitch only shifted): synthetic version with f_0 shifted up toward values typical of women and spectrum envelope unmodified from measured values. (4) MEO (men, envelope only shifted): synthetic version with the envelope shifted up toward values typical of women and f_0 unmodified from measured values. The same approach was taken with the signals spoken by women to create four additional conditions, but with f_0 and/or envelope shifted down in frequency toward values typical of men: (1) WUS (women, unmodified synthesis), (2) WPE (women, pitch and envelope shifted), (3) WPO (women, pitch only shifted), and (4) WEO (women, envelope only shifted).

Source-filter synthesizer. All test signals were generated with the spectral envelope synthesizer (SES), which is described in detail in Hillenbrand, Houde, and Gayvert (2006). Figure 3 shows a source

Table 1
Accuracy in Identifying Speaker Sex Using a Discriminant Classifier

		Absolute Frequencies		Normalized Frequencies	
	Parameters	PC	*SD*	PC	*SD*
f_0 only	f_0	95.8	0.4	96.3	0.5
Formants only	F_1	59.0	1.4	81.3	1.0
	F_2	52.3	2.2	81.5	1.8
	F_3	78.2	1.5	86.1	1.0
	F_1, F_2	70.2	1.8	87.8	1.0
	F_1–F_3	79.2	1.8	91.9	1.0
	Geometric mean F_1–F_3	68.8	2.0	92.4	0.9
Formants and f_0	f_0, F_1	98.3	0.5	97.2	0.4
	f_0, F_1–F_2	96.4	0.4	97.8	0.5
	f_0, F_1–F_3	96.7	0.4	98.4	0.4

Note—Percent correct (PC) figures are averages across 30 random split-half tests.

signal and frequency response curve that would be used to generate a monotone, sustained, phonated /ɑ/. The source signal here is a sequence of single-sample pulses whose period is determined by the measured instantaneous fundamental period of the signal that is being reconstructed. A whispered vowel (or any other unvoiced segment) can be synthesized by replacing the periodic pulse train that is shown in Figure 3 with a sequence of single-sample pulses whose amplitudes are either zero or nonzero, with a probability of .5 at each sample point. Mixed-source signals of the kinds that would be needed to synthesize speech sounds such as breathy vowels or voiced fricatives (i.e., consisting of both periodic and aperiodic components) can be generated simply by adding periodic and random pulse sequences with any desired voiced/unvoiced mixing ratio. f_0 and degree of periodicity (needed to set the amplitude ratio of periodic to aperiodic pulses) are measured with a cepstrum-like pitch tracker, described in Hillenbrand et al. (2006).

The filter function shown in Figure 3 is simply the spectrum envelope of the speech signal, which, in turn, is used directly to define a finite impulse response filter to modify the spectral shape of the flat-spectrum source signal. The spectral envelope is estimated using a method that we have called the *harmonic envelope* (see Paul, 1981, for a related method). As is shown in panel A of Figure 4, this method simply involves linearly interpolating between the harmonics of a narrow band (512 points, 32 msec) Fourier spectrum, using the same f_0 measurements that are used to create the source signal. Despite our use of the term *harmonic envelope*, the method makes no distinction between periodic and aperiodic speech segments. f_0 is measured for all speech frames, regardless of the degree of periodicity, and these estimates are used to define the envelope for unvoiced and marginally periodic segments using the method just described, in spite of the fact that the peaks will often not correspond to harmonics. Figure 4B shows an example of a harmonic envelope for a spectral slice taken from a stop release burst for a [k^{h}]. Once the source signal and frame-by-frame sequence of impulse responses have been computed from the spectral envelopes, synthesis is simply a matter of convolving the source signal with the time-varying impulse response.

Earlier work showed that sentence intelligibility for the SES, evaluated using utterances drawn from the same subset of the TIMIT database that was used in the present study, was excellent (96.9%, Hillenbrand et al., 2006).

Modification of f_0 and envelope. For the conditions in which the fundamental frequencies of men were shifted toward values typical of women (MPE and MPO), all f_0 values were scaled by 1.7041, the average ratio of adult female and male fundamental frequencies in the Peterson and Barney (1952) data. For the conditions in which the spectral envelopes of men were shifted toward values typical of women (MPE and MEO), the spectral envelopes were shifted up in frequency by a factor of 1.168, the average ratio of adult female and male formant frequencies (F_1–F_3) in the Peterson and Barney data. Inverses of these values were used to shift the f_0 values and envelopes of women down in the direction of those for men.

Subjects and Procedure. Twenty-one students majoring in speech–language pathology served as listeners. Subjects passed a 25-$\mathrm{dB}_{\mathrm{HL}}$ hearing screening at octave frequencies between 125 and 4000 Hz. On each trial, listeners were asked to choose between two buttons on a computer screen to judge whether the sentence was spo-

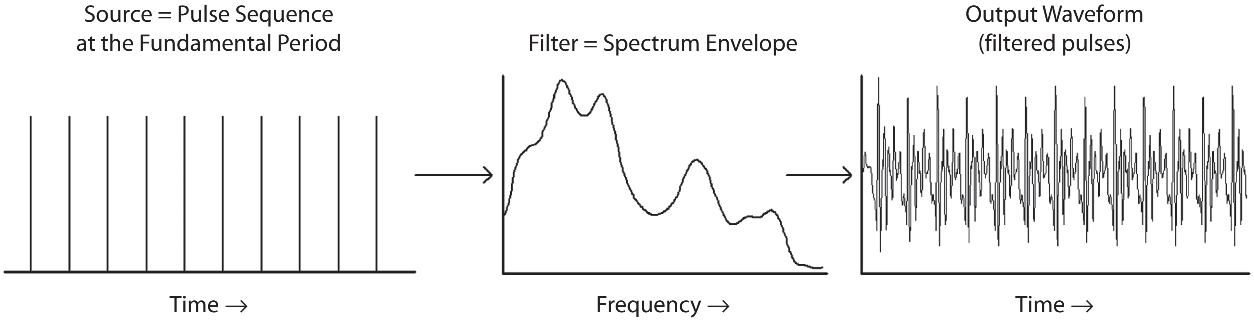

Figure 3. Illustration of the synthesis of a sustained vowel using the spectral envelope synthesizer. The source is a sequence of single-sample, spectrally white pulses spaced at the fundamental period. The filter is the spectrum envelope measured from the speech signal being synthesized. A whispered vowel can be synthesized simply by replacing the periodic source signal with a sequence of pulses spaced at random intervals.

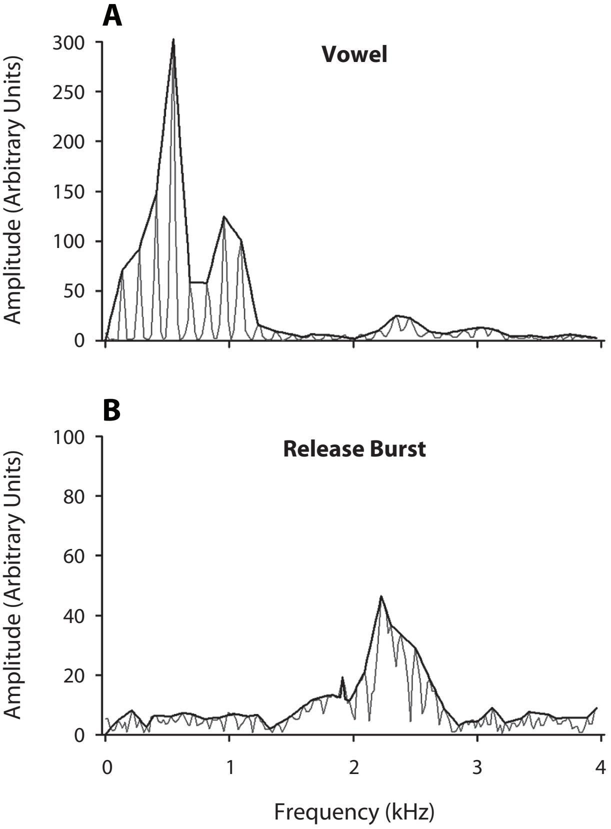

Figure 4. The harmonic envelope for a vowel (A) and for a consonant release burst (B). Spectra shown up to 4 kHz only.

ken by a man or a woman. Listeners were then asked to estimate their confidence in their judgment by choosing among buttons labeled 1 (*lowest confidence*) through 5 (*highest confidence*). The listening task consisted of 200 trials: 50 sentences (25 spoken by men and 25 by women), each presented under four conditions (unmodified synthesis, f_0 shifted, envelope shifted, f_0 and envelope shifted). The 200 sentences were presented in a single random order (i.e., not blocked by condition), scrambled separately for each listener. The stimuli were low-pass filtered at 7.2 kHz, amplified, and delivered free field in a quiet room over a single loudspeaker (Paradigm Titan v.3) positioned about 1 m from the listener's head at a level averaging about 75 dBA for the stressed syllables.

Results

Results of Experiment 1 are shown in Figure 5. Panel A shows percent identification as men (dark bars) or women (shaded bars) for each of the conditions. It can be seen that speaker sex is conveyed almost perfectly in the two conditions in which f_0 and envelope were unmodified from their original measured values (MUS, WUS). This shows that the SES preserves information from the original signal that is sufficient to convey speaker sex. Of course, even these very high identification rates should not be taken to mean that the synthesizer conveys all information that is relevant to speaker sex, only that the information that is conveyed is easily sufficient. Perceived speaker sex was altered for the majority, but by no means all, of the stimuli for which both f_0 and formants were shifted. Approximately 82% of signals in the MPE condition (spoken by men but with f_0 and envelope shifted up) were heard as women, and a nearly identical percentage of signals in the WPE condition (spoken by women but with f_0 and envelope shifted down) were heard as men. Shifting either one of these features alone was far less effective in altering perceived speaker sex than was shifting both. Shifting f_0 alone for utterances originally spoken by men (MPO) resulted in ~34% of the signals being heard as women, while shifting only f_0 for utterances originally spoken by women (WPO) resulted in just 19% of the signals being heard as men. Modifying envelope alone resulted in even fewer shifts in perceived talker sex, with only 19% of the MEO signals being heard as women and just 12% of the WEO signals being heard as men. It can also be seen that, as measured by absolute identification, the individual effects of f_0 and envelope do not sum to the combined effects of the two features; for example, the sum of the identifications as women of MPO (34%) and MEO (19%) is much less than the percentage of identifications as women under the MPE condition. This underadditivity is even more striking for the W conditions, with the WPO and WEO conditions summing to just 31%, far less than the 82% for the WPE condition.

A one-way repeated measures ANOVA comparing the arcsine-transformed means shown in panel A of Figure 5 showed a highly significant effect for condition [$F(7,140) = 132.6$, $p < .0001$]. Results of Bonferroni planned comparisons included the following: (1) The MUS and WUS conditions did not differ from one another but were each reliably different from all conditions in which f_0, envelope, or f_0 and envelope were shifted; (2) within the group of M conditions (i.e., MPE, MPO, MEO) and within the group of W conditions (i.e., WPE, WPO, WEO), all pairwise comparisons were significantly different; and (3) in comparing corresponding conditions for male versus female talkers (e.g., MPE vs. WPE, MPO vs. WPO), only the MPO versus WPO difference was significant.

Panel B of Figure 5 shows mean confidence ratings for each condition. At roughly 4.6 on a 5-point scale, confidence ratings were quite high for the two unmodified synthesis conditions and noticeably lower for all of the frequency-shifted conditions (i.e., pitch and envelope shifted [PE], pitch shifted only [PO], and envelope shifted only [EO]). Despite rather large differences in identification rates across the individual frequency-shifted conditions (e.g., MPE vs. MPO vs. MEO, etc.), differences in confidence ratings across these conditions were not especially large. A one-way repeated measures ANOVA comparing the mean confidence ratings shown in panel B of Figure 5 showed a highly significant effect for condition [$F(7,140) = 50.5$, $p < .0001$]. Results of Bonferroni planned comparisons included the following: (1) The MUS and WUS conditions did not differ from one another but were each reliably different from all of the frequency-shifted conditions; (2) within the group of M conditions, only the MPE–MEO comparison did not reach significance; (3) there were no significant differences within the group of W conditions; and (4) only the MPO–WPO

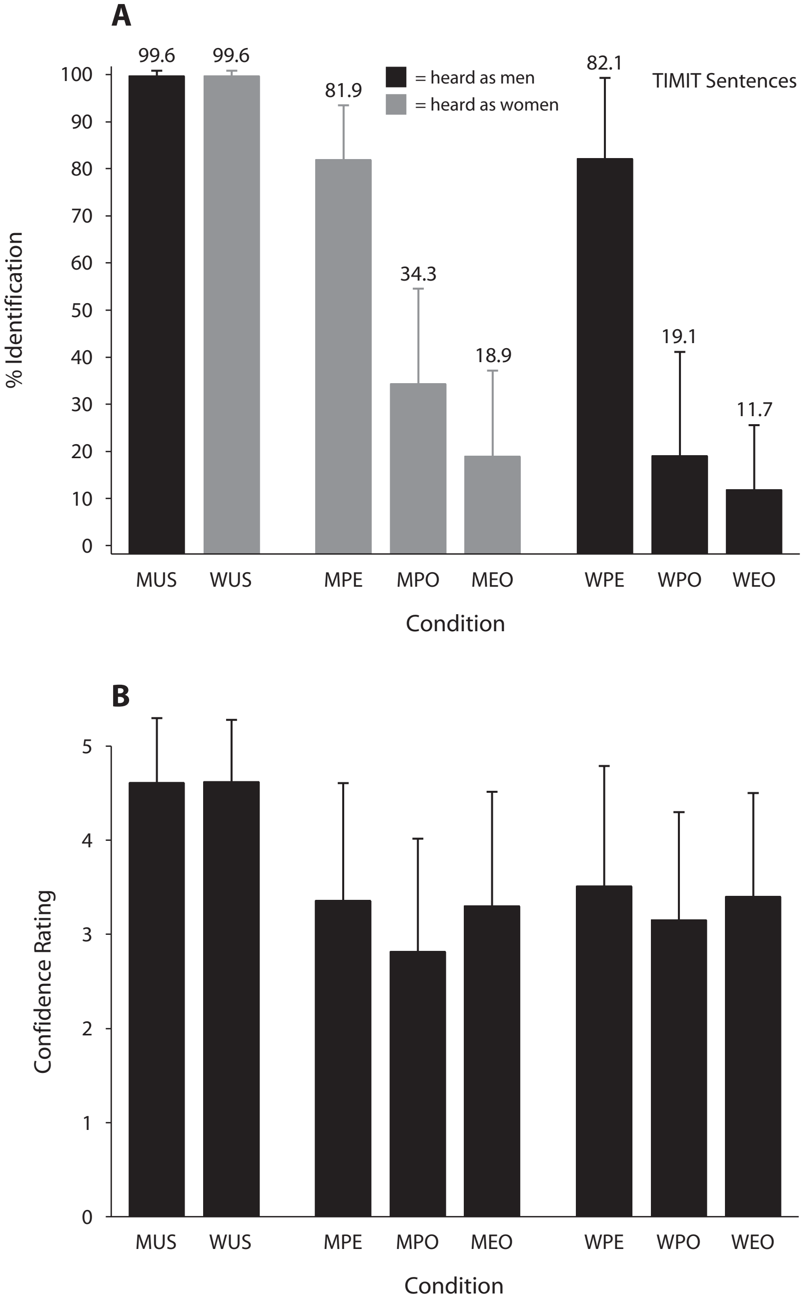

Figure 5. (A) Percent identification as men (dark bars) or women (shaded bars) for each condition. (B) Confidence ratings for the same conditions. Error bars in both panels indicate one standard deviation. Conditions: MUS = men, unmodified synthesis; WUS = women, unmodified synthesis; MPE = men, pitch and envelope shifted; MPO = men, pitch only shifted, MEO = men, envelope only shifted; WPE, WPO, WEO: corresponding conditions for adult female talkers.

comparison reached significance in comparisons of corresponding conditions for male versus female talkers (MPE–WPE, MPO–WPO, MEO–WEO). In general, the strongest impression created by the confidence ratings is that listeners found the frequency-shifted signals to be more ambiguous than the unmodified conditions, even for combined f_0 and envelope shifts, which changed perceived talker sex fairly strongly.

Discussion

The main findings from Experiment 1 include the following: (1) Talker sex was conveyed almost perfectly by the SES when driven by the original measured values. (2) Perceived talker sex shifted rather strongly from male to female when both f_0 and envelope were shifted up, and to a nearly equal degree from female to male when both f_0 and envelope were shifted down, although in both cases a substantial minority of signals (~18% in both cases) retained their original sex identity even with shifts in both parameters. (3) On a substantial majority of trials, single-parameter shifts were ineffective in changing perceived talker sex. (4) By itself, f_0 was more effective in shifting perceived talker sex than was envelope, although the difference reached significance for men only. (5) As measured by absolute identification, the individual effects of f_0 and envelope did not sum to equal the combined effects of the two features. (6) In the absolute labeling data, there were no significant departures from symmetry across male and female talkers (i.e., upward frequency shifts for men produced as many changes in perceived talker sex as downward frequency shifts produced for women). (7) Confidence ratings indicated that listeners found the frequency-shifted signals to be more ambiguous than the unmodified signals, even for the f_0 and envelope manipulation, which produced shifts in perceived talker sex on a substantial majority of trials.

The most obvious conclusion to be drawn from the ~18% of trials in which utterances retained their original perceived talker sex despite substantial shifts in both f_0 and envelope is that there is more to the perception of speaker sex than f_0 and formants. Among the many candidates for residual cues to speaker sex are possible male–female differences in the prosodic characteristics of utterances. Any male–female differences in prosodic features would be expected to be conveyed well by the sentences used in Experiment 1. These features, however, are not present to any interesting degree in single-syllable utterances spoken in citation form. To gain some initial insight into the role that might be played by prosodic features in the perception of speaker sex, we ran a second experiment consisting of a replication of Experiment 1 using /hVd/ utterances, the great majority of which were spoken with a simple falling pitch contour. If syllables are found to change perceived gender with frequency shifts more readily than sentences (and we will report some evidence for this below), it would suggest some role for prosodic features in distinguishing men and women.

There is a second reason for repeating Experiment 1 with syllables. Returning again to the ~18% of trials in which utterances retained their original perceived talker sex despite shifts in both f_0 and envelope, at least one other explanation presents itself in addition to demonstrating the role of residual cues to speaker sex. It is possible that some of this ~18% figure is a simple side effect of the particular approach that was taken to frequency shifting. Recall that all stimuli were shifted by the same scale factors, which were derived from average f_0 and formant ratios between men and women. Consider, for example, a man with atypically low formants and/or f_0. It is possible that the fixed upward frequency scale factors that were used are simply not large enough to shift utterances such as these to values that are typical of women. The same, of course, might be said of a woman with atypically high f_0 and/or formant values. As is explained below, this can be sorted out after the fact by measuring correlations between perceived talker sex and the f_0 and formant characteristics of the original utterances. Although this possibility could be evaluated with acoustic analyses of the TIMIT sentences from Experiment 1, the problem is greatly simplified, especially for formant measurements, by using syllables whose acoustic properties can be characterized with a high degree of confidence.

EXPERIMENT 2

Method

/hVd/ syllables spoken by 25 men and 25 women were drawn from the 16-kHz digital recordings made by Hillenbrand et al. (1995). Vowels at the corners of the quadrilateral (/i/, /æ/, /ɑ/, /u/) were used. Within those constraints, signals were drawn at random from the larger database. The utterances selected in this way were spoken by 23 different men and 20 different women, but with uneven numbers of tokens across the 43 talkers. The synthesis methods described for Experiment 1 were used to create the same stimulus conditions as were used in Experiment 1 (i.e., MUS, WUS, MPE, etc.), resulting in 200 test signals. Earlier work (Hillenbrand et al., 2006) showed that vowels synthesized with the SES were as intelligible as the naturally spoken signals on which they were based.

Subjects and Procedure. A separate group of 24 normal-hearing students enrolled in an introductory phonetics course served as listeners. Procedures for randomizing trials, delivering stimuli, and collecting responses were identical to those of Experiment 1.

Results

The hatched bars in panel A of Figure 6 show percent identification as men (dark bars) or women (shaded bars) for each condition. For comparison, findings from the sentences used in Experiment 1 are reproduced in the solid bars. Some features of the syllable data are quite similar to the corresponding data from Experiment 1. In particular, identification rates are quite high for the two unmodified synthesis conditions, and the majority of the f_0 and envelope-shifted signals changed perceived talker sex in the expected direction. However, in the case of the WPE signals, the shift was roughly 10 percentage points higher than for the corresponding condition in Experiment 1. The f_0-only conditions also produced substantially more identification shifts than in Experiment 1, by almost 22 percentage points for men and roughly 49 percentage points for women. Envelope shifts alone were largely ineffective in altering perceived talker sex for both sentences and syllables, although for the syllables the number of identification shifts for the women was nearly double that for the sentences.

Focusing first on the new results from Experiment 2, a one-way ANOVA on the arcsine-transformed percent identification values in the hatched bars in Figure 6 showed a highly significant effect for condition [$F(7,161) = 194.7$, $p < .0001$]. Bonferroni planned comparisons included the following. (1) The MUS and WUS conditions did not differ from one another but were each reliably different from all six frequency-shifted conditions. (2) Within the group of M conditions (i.e., MPE, MPO, MEO), all pairwise comparisons were significantly different, as were all pairwise comparisons within the group of W conditions. (3) Comparing corresponding conditions for male versus female talkers, the greater percentage of shifts in

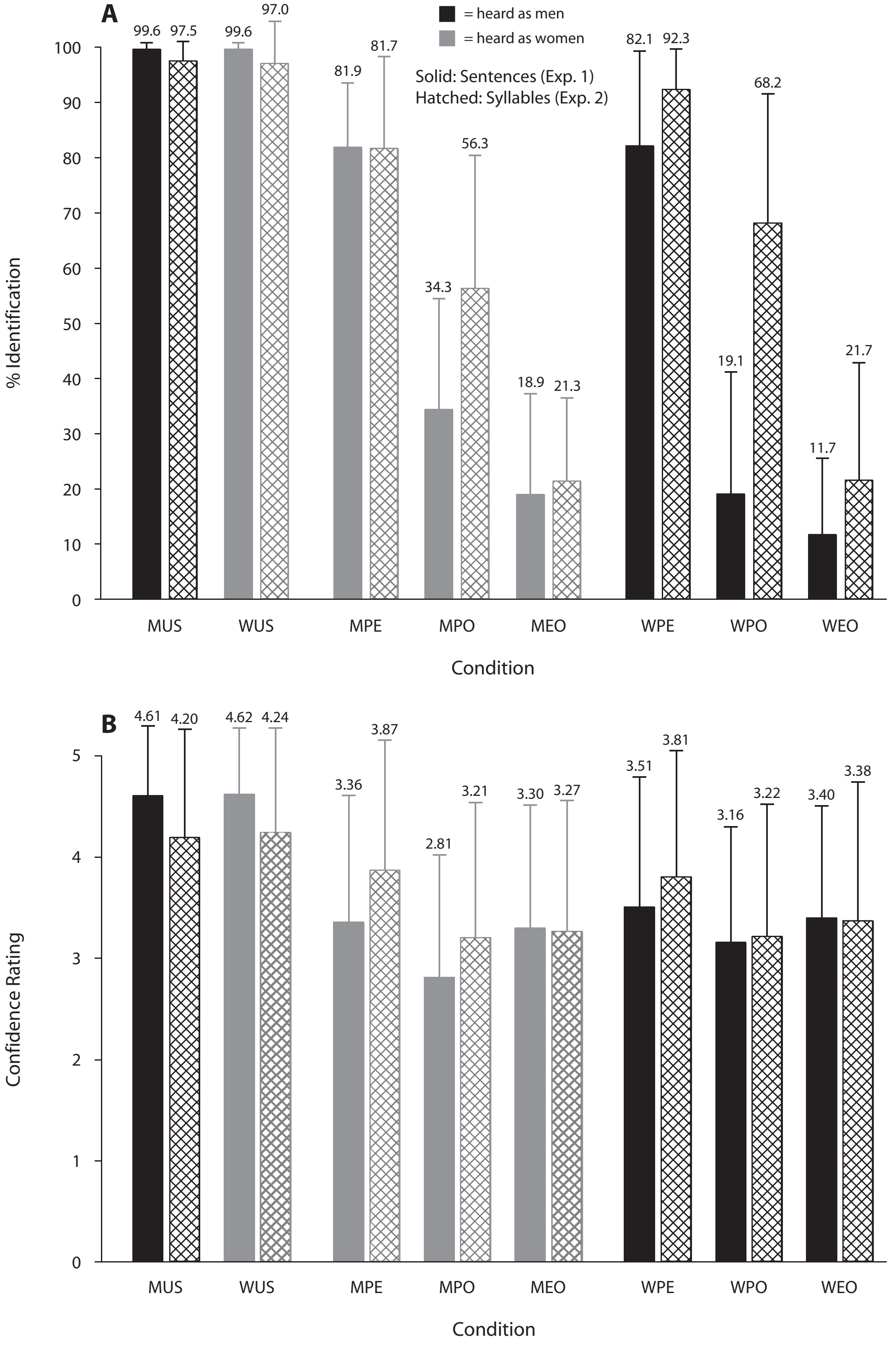

Figure 6. (A) Percent identification as men (dark bars) or women (shaded bars) for each condition. (B) Confidence ratings for the same conditions. Solid bars indicate sentences (Experiment 1); hatched bars indicate syllables (Experiment 2). Error bars in both panels indicate one standard deviation. Conditions: MUS = men, unmodified synthesis; WUS = women, unmodified synthesis; MPE = men, pitch and envelope shifted; MPO = men, pitch only shifted, MEO = men, envelope only shifted; WPE, WPO, WEO: corresponding conditions for adult female talkers.

perceived gender for women than for men in the PE and PO conditions reached significance.

A one-way ANOVA comparing the mean confidence ratings shown in panel B of Figure 6 showed a significant effect for condition [$F(7,161) = 39.8$, $p < .0001$]. Planned comparison results included the following. (1) The MUS and WUS conditions did not differ from one another but were each different from all of the frequency-shifted conditions. (2) Within the group of M conditions, only the MPO–MEO comparison did not reach significance. (3) Similarly, only the WPO–WEO comparison failed to reach significance within the group of W conditions. (4) There were no significant differences in comparisons of corresponding conditions for male versus female talkers.

To examine differences in absolute identification between the sentences of Experiment 1 and the syllables of Experiment 2, a two-way ANOVA was calculated using the arcsine-transformed percent identification values whose means are shown in Figure 6. The analysis showed highly significant effects for condition [$F(7,42) = 194.7$, $p < .0001$], speech material [$F(1,42) = 44.0$, $p < .0001$], and—the only feature of this analysis that was not evaluated by the previous ANOVAs—a significant interaction [$F(7,38) = 13.9$, $p < .0001$]. The source of the interaction, which can be readily seen by examining panel A of Figure 6, is the larger number of shifts in perceived gender for the syllables. Bonferroni planned comparisons showed significantly more shifts in perceived gender for syllables than for sentences for the MPO, WPE, and WPO conditions.

The same type of analysis for confidence ratings showed an effect for condition [$F(7,42) = 83.9$, $p < .0001$], no main effect for speech material [$F(1,42) = 0.03$, n.s.], and a significant interaction [$F(7,38) = 7.8$, $p < .0001$]. Although the interaction was highly significant, the only statistically reliable planned comparison between confidence ratings for corresponding conditions for sentences versus syllables was for the MUS condition, which showed higher confidence ratings for the sentences.

Discussion

Findings from the /hVd/ syllables used in Experiment 2 are generally similar to those from the sentences used in Experiment 1, but there are some differences of interest. The most prominent is that shifts in f_0 alone were much more likely to produce a change in perceived speaker sex for syllables than for sentences. This difference was especially striking for the f_0-shifted syllables originally spoken by women, which were nearly four times more likely to be heard as men than were the f_0-shifted sentences. The f_0- and envelope-shifted syllables spoken by women were also about 10 percentage points more likely to be heard as men than were similarly shifted sentences, although this kind of difference was not observed for the utterances originally spoken by men. In general, and despite asymmetries between utterances spoken by men and women, syllables shifted perceived speaker sex more readily than did sentences. This kind of difference between Experiments 1 and 2 would suggest that residual cues to speaker sex are conveyed to a greater degree in sentence-length material than in isolated syllables. Possible explanations are explored in the General Discussion section.

CORRELATIONS WITH ACOUSTIC MEASUREMENTS

As noted above, there are two different and nonexclusive explanations that might account for the observation that a significant minority of signals in both experiments failed to shift perceived gender, in spite of sizable shifts in both f_0 and envelope. Although this finding almost certainly reveals, in part, the importance of cues to speaker sex other than f_0 and envelope, some signals might have retained their original sex identity simply because the frequency shifts were not large enough for men with atypically low f_0 and/or formants, or for women with atypically high f_0 and/or formants. This can be evaluated by measuring correlations between perceived gender and the acoustic characteristics of the original utterances. The syllables are particularly well suited to this kind of analysis, since the acoustic properties of these utterances, particularly the formants, are much easier to characterize than those of sentences. And although no simple method presents itself for normalizing formant frequencies for the TIMIT sentences,[3] f_0 measurements from these utterances can be correlated with the listener data.

Method

For the TIMIT sentences, average f_0 was measured from hand-edited versions of the pitch contours that were used to drive the SES. Editing consisted primarily of deleting f_0 measurements during unvoiced and marginally periodic regions of the sentences (see H95 for a description of the editing tool). For the syllables, hand-edited measures of f_0 and F_1–F_3, sampled at the steadiest portion of the vowel, were taken from H95. Formant frequencies were normalized using the standard-score method described above. In light of the (disputed) suggestion of increased prosodic variability in women, measures were also made of f_0 variability for the TIMIT sentences. Prosodic variability, relevant for the sentences only, was measured as the *SD* of f_0 in semitones, relative to an arbitrarily selected base frequency of 50 Hz. All measurements were made from the original recordings.

Results

Results are reported in Table 2, which shows, for the PE conditions only, correlation coefficients relating the percentage of judgments that did *not* reflect a shift in perceived gender with (1) mean f_0, (2) mean normalized F_1–F_3, and (3) f_0 variability. Results are shown for the TIMIT sentences of Experiment 1 and the /hVd/ syllables of Experiment 2. Measurements of normalized formants are available for the syllables only, and correlations with prosodic variability, relevant for the sentences only, are not reported for the /hVd/ syllables. Only two of the correlations are reliably different from zero: (1) for the TIMIT sentences, and for utterances spoken by women only, a tendency for utterances spoken at higher f_0 to retain their female gender in spite of downward shifts in f_0 and envelope ($r = .60$), and (2) for the syllables, and

Table 2
For the Pitch and Envelope Conditions, Correlations Relating the Percentage of Utterances That Did Not Shift Perceived Gender With Selected Acoustic Measures

	TIMIT Sentences		/hVd/ Syllables	
	Men	Women	Men	Women
Mean f_0	−.15	**.60**	**−.49**	.04
Mean normalized F_1–F_3	–	–	−.24	.32
f_0 variability (*SD* in semitones)	−.09	−.25	–	–

Note—Results are shown for the TIMIT sentences of Experiment 1 and the /hVd/ syllables of Experiment 2. Measurements of normalized formants are available for the syllables only, and correlations with prosodic variability, relevant for the sentences only, are not reported for the /hVd/ syllables. The two correlations that are significant at .05 or less are shown in bold.

for utterances spoken by men only, a tendency for utterances spoken at lower f_0 to retain their male gender in spite of upward shifts in f_0 and envelope ($r = -.49$). For the women's syllables, the correlation with normalized formants fell just short of significance. Correlations with intonation variability for the TIMIT sentences were weak and statistically indistinguishable from zero. The measurement data, in fact, showed slightly (but not significantly) *greater* f_0 variability for the men than for the women [2.25 vs. 2.21 semitones, respectively; $t(48) = 0.61$, n.s.]. The main conclusion to be drawn from these results is that factors other than the average f_0, f_0 variability, and the normalized formant frequencies of the original speech samples account for *most* of the utterances that retained their original sex identity despite pitch and envelope shifting.

GENERAL DISCUSSION

The purpose of this study was to measure the contributions of f_0 and formants in cuing the distinction between men's and women's voices. A preliminary pattern recognition study using f_0 and formants measured from /hVd/ syllables showed that f_0 alone (~96%) distinguishes speaker sex somewhat better than do normalized formants alone (~92%), and that the combination of f_0 and normalized formants distinguishes speaker sex nearly as well as human listeners do. Raw formant frequencies, however, distinguish speaker sex rather poorly, suggesting that the use of formant information to recognize speaker sex may depend on the recognition of vowel identity. This suggests a potentially problematic "chicken and egg" mutual interdependency, such that the recognition of vowel identity depends on scaling information of the sort that distinguishes speaker sex, yet the use of that scaling information to distinguish men's and women's voices may itself depend on knowledge about vowel identity. There is, in fact, some experimental evidence for this mutual interdependency in human listeners. In a deceptively simple study, Eklund and Traunmüller (1997) found that the error rate for identifying whispered Swedish vowels spoken by men and women was nearly five times greater on trials in which the sex of the speaker was misidentified than on the remaining trials. This suggests that information specifying speaker sex (or, alternatively, some acoustic correlate of vocal tract length) is used in vowel recognition, a conclusion that would not surprise any advocate of vocal tract normalization. What is most striking about the Eklund and Traunmüller findings, however, is that the reverse also appears to hold: The error rate for identifying speaker sex was more than four times greater on trials in which the vowel was misidentified than on the remaining trials. (These interdependencies were not observed for phonated vowels, but the error rates for both vowel and sex identification were quite low for these signals.) For whispered vowels—signals for which the evaluation of formant/envelope information is especially critical—it seems difficult to avoid the conclusion that speaker sex identification depends on vowel identification *and* that vowel identification depends on speaker sex identification (or on the underlying scaling information). Perhaps we suffer from a lack of imagination due to overly intimate experience with sequential computing algorithms, but it is not immediately obvious how a recognition scheme of this kind might be implemented.

In listening experiments using both synthesized sentences and isolated syllables, the majority of signals that were synthesized with shifts in the appropriate direction of both f_0 and formants changed perceived talker sex. This finding is consistent with the commonly held view that these two features, which have the most direct and obvious connection with anatomical differences in the speech structures between men and women, are the major determinants of the speaker-sex percept among adult talkers. There was also reasonably clear evidence that f_0 plays a somewhat larger role than envelope does: For both sentences and syllables, shifts in f_0 alone produced more changes in perceived talker sex than did shifts in envelope alone (a finding that is consistent with statistical pattern recognition results showing a modest advantage in classification accuracy for f_0 alone as compared with normalized formants alone). Frequency shifts in either parameter alone, however, were largely ineffective in altering perceived sex; that is, when a single parameter was altered, listeners usually heard the sex of the original talker. Closely related to this is the finding, primarily for the more realistic case of connected speech, that the individual effects of f_0 and envelope scaling do not sum to the combined effects of the two features (i.e., the combined effects of f_0 and envelope scaling are much greater than the sum of their individual effects). We take this "underachievement" of single-parameter frequency shifts as an indication that residual cues to speaker sex (in the present context, any cues to speaker sex that are preserved in the face of frequency scaling) are given greater weight by listeners when the primary cues of f_0 and formants are ambiguous. In an extensive series of experiments on the effects of f_0 and formant scaling on speech intelligibility, Assmann and colleagues have shown quite clearly that listeners have well-developed internalized knowledge about the relationships that typically prevail between f_0 and formant frequencies (e.g., Assmann & Nearey, 2003a, 2003b, 2007; Assmann, Nearey, & Scott, 2002). Their findings show that speech intelligibility suffers when these expectations are violated. Results from the single-cue conditions in the present study suggest that listeners will tend to rely more heavily on cues to speaker sex other than f_0 and formant/envelope scaling when these expectations are violated.

There are other clear indications in the findings that there is more to the speaker-sex percept than just the scaling of f_0 and formants, the most straightforward being the significant minority of signals that retained their original sex identity after large shifts in both f_0 and envelope. Also important are the significantly higher confidence ratings for the unshifted utterances as compared with all conditions involving frequency shifts, including shifts in both parameters. Although f_0 and formants clearly play the dominant role in distinguishing the voices of men and women, these cues are clearly not quite sufficient. Listeners do not find frequency-shifted men's voices to be entirely convincing as women, and, to a nearly equal degree, they do not find frequency-shifted women's voices to be entirely convincing as men. A similar conclusion was reached by Assmann et al. (2006), although, in that study, the signals were modeled on the utterances of just two men and two women (although with a very broad sampling of frequency shifts). Further, Klatt and Klatt's (1990) dissatisfaction with the authenticity of formant-synthesized women's voices produced by scaling the f_0 and formant values derived from analyses of male speech was the primary motivation behind their insightful series of experiments on male–female voice quality differences.

If f_0 and formant scaling are not sufficient to signal speaker sex, what are these residual cues to speaker sex? Our study was not designed to address this question, but the results may provide a few hints. The major difference between Experiments 1 and 2, which were essentially identical except for speech material, was that sentences were more resistant than syllables to changes in perceived speaker sex. A similar difference can be seen by comparing the results of Smith et al. (2007), who reported little evidence for residual cues to speaker sex in their study of isolated vowels, with those of Assmann et al. (2006), who reported clear evidence for them in their study of sentences. Taken together, these findings suggest that residual cues to speaker sex are more readily conveyed in sentences than in isolated syllables. This might lead one to think of the possibility of male–female differences in prosodic features, which would not be conveyed well in isolated vowels or syllables. The single measure that we made of f_0 variability showed no differences between men and women, but of course there is much more to the melody and rhythm of speech than would be captured by this simple measure. This would be a worthwhile avenue for further work on this problem.

The voice-quality differences between men and women that were studied by Klatt and Klatt (1990) are probably the most extensively studied of the proposed secondary cues to speaker sex. The authors found that, on average, women tend to have somewhat breathier voices than men do. In principle, the source-filter synthesizers used in the present study, and in the earlier frequency-shifting studies, are able to convey the relevant underlying acoustic features (primarily aspiration noise and the spectral tilt of the harmonic component). It remains to be determined whether this information is, in fact, conveyed with sufficient precision by these synthesis methods, and the perceptual importance of this cue in relation to the major cues of f_0 and formants is as yet unknown.

The method used here and in the earlier frequency-shifting studies involved shifting the envelope as a whole, implicitly assuming a uniform scaling relationship between the formants of men and women. It is well-known that the scale factors that relate the formants of men and women are not quite uniform, with scale factors varying across formants and, especially, across vowels (e.g., Fant, 1975; Traunmüller, 1984, 1988). These nonuniform scaling relationships have been extensively studied, but the primary focus of this work has been on the implications of this nonuniformity for phonetic recognition. It is possible that sex-dependent and vowel-specific scaling relationships (preserved with uniform frequency scaling) play some role in the perception of speaker sex. Our speculation, though, is that this kind of cue did not play a dominant role in the present study. It seems reasonable to assume that this kind of cue would be easier to apprehend in syllables since there listeners can focus more closely on detailed vowel quality, while in sentences they have much more to attend to. Yet we found that syllables were more rather than less likely to change speaker sex with frequency scaling. Sex-dependent, vowel-specific scaling relationships, however, remain a viable candidate.

Our findings do not provide consistent support for the Owren et al. (2007) hypothesis of a male advantage in the perception of speaker sex. The suggestion that "the presence of critical features of 'maleness' virtually guarantees that the talker is an adult male, [but] their absence does not unequivocally imply that the talker is an adult female" (p. 930) would seem to predict that downward shifts in f_0 and/or formants (i.e., producing the "critical features of 'maleness'") would be more effective in eliciting a change in perceived speaker sex than would upward frequency shifts. No such asymmetry was found in our sentence data: Upward shifts in both parameters were just as likely to elicit changes in perceived talker sex as were downward shifts. Furthermore, for single-parameter manipulations in both f_0 and envelope, upward shifts were somewhat *more* likely to elicit changes in perceived talker sex than were downward shifts (Figure 5). The syllable data from Experiment 2, however, are more compatible with the Owren et al. hypothesis: Downward shifts in f_0 and formants produced more changes in perceived speaker sex than did upward shifts. The same was true for shifts in f_0 alone, although no asymmetry was seen for shifts in envelope alone (Figure 6). Considering results from both sentences and syllables collectively, we found no consistent evidence supporting the view that listeners tend to hear voices with conflicting or ambiguous acoustic properties as male, or for the idea that low values of f_0 or formants are in some sense more readily interpreted as signs of masculinity than high values are interpreted as signs of femininity. There is, of course, no reason to doubt male–female performance asymmetries, which have been reported in several studies (e.g., Bennett & Montero-Diaz, 1982; Coleman, 1971, 1976; Lass et al., 1976; Owren et al., 2007), but it may be that there are aspects of this phenomenon that are not yet well understood.

A final point worth noting about these findings, along with other results confirming a significant role for the envelope in the perception of speaker sex, concerns the fundamental question about the psychological reality of formants in speech perception. The concept of formants is, of course, pervasive in the speech perception literature, and it is uncommon to get very far into nearly any discussion of phonetic perception without some reference being made to formants. Despite the widespread use of the formant pattern as an explanatory concept in speech perception, and the numerous virtues of formant representations, the idea is not without some troublesome problems, which have been noted by a number of investigators. Briefly, these problems include the following: (1) the *determinacy problem*, as Bladon (1982) has called it, which is the commonplace idea that tracking formants in natural speech is a difficult and, as yet, unresolved problem; (2) the straightforward observation that perceptual confusions made by human listeners nearly always involve speech sounds that are phonetically quite similar, a pattern that is difficult to reconcile with an underlying formant tracking process that is susceptible to gross errors that occur when formants either split or merge[4] (Klatt, 1982; see also Ito, Tsuchida, & Yano, 2001); (3) evidence showing that spectral details other than formant frequencies can affect phonetic quality (e.g., Bladon, 1982; Chistovich & Lublinskaja, 1979; Hillenbrand & Nearey, 1999). Partially in response to these problems, a number of investigators have argued that phonetic recognition is mediated by mental computations of similarities and differences in the gross shape of the spectrum rather than by formant frequencies (e.g., Bladon & Lindblom, 1981; de Cheveigné & Kawahara, 1999; Hillenbrand & Houde, 2003; Zahorian & Jagharghi, 1993). The issues discussed above revolve around the question of formants versus gross spectral shape in the perception of phonetic quality rather than talker sex, but the same question can be asked about either issue. Whether implicit or explicit, the universal assumption is that the acoustic consequences of vocal tract length differences between men and women (or children and adults, or children of different ages) are conveyed to listeners by some measure derived from the formant frequency pattern. As was noted above, whole-spectrum models of vowel recognition have been developed that are capable of classifying vowels on the basis of the gross shape of the spectrum. Can similar solutions be found for the speaker-sex problem; specifically, is there a way to explain how listeners perceive the "envelope height" differences between like vowels spoken by men and women, for example, without appealing to formants? Although by no means unsolvable, this does not strike us as a simple problem. We offer no solution here but would recommend this problem as a worthy topic for further work in this area.

SUMMARY

The main findings of these experiments were as follows. (1) A statistical pattern classifier differentiated syllables spoken by men and women accurately on the basis of either f_0 or formants alone, although somewhat more accurately for f_0 alone. (2) Statistical separation was more accurate when both f_0 and formants were used. (3) Distinguishing men's and women's voices based on formants was far better when a representation was used that took vowel identity into account. (4) A substantial majority of sentences (~82%) changed perceived talker sex when both f_0 and envelope were frequency shifted in the appropriate direction, with the changes in perceived sex occurring about equally often from men to women as vice versa. (5) An important minority of signals retained their original sex identity and evoked lower confidence ratings, despite substantial shifts in both f_0 and envelope. (6) Confidence ratings for all frequency-shifted conditions were lower than those for the conditions with unmodified synthesis parameters. (7) Single-parameter shifts in f_0 and, especially, in envelope were largely ineffective in altering perceived sex. (8) A second experiment using isolated syllables produced results that were largely similar to those for sentences, with the important exception that syllables shifted perceived sex more readily than did sentences with changes in f_0 alone or in f_0 and envelope together, although this speech-material effect was uneven across male and female talkers.

AUTHOR NOTE

This work was supported by a grant from the National Institutes of Health (R01-DC01661) to Western Michigan University. We are grateful to Steve Tasko, Greg Flamme, Peter Assmann, and an anonymous reviewer for helpful comments on an earlier draft. Address correspondence to J. M. Hillenbrand, Speech Pathology and Audiology, MS5355, Western Michigan University, Kalamazoo, MI 49008 (e-mail: james.hillenbrand@wmich.edu).

REFERENCES

Ainsworth, S. (1975). Intrinsic and extrinsic factors in vowel judgments. In G. Fant & M. A. A. Tatham (Eds.), *Auditory analysis and the perception of speech* (pp. 103-113). London: Academic Press.

Assmann, P. F., & Nearey, T. M. (2003a). Frequency shifts and vowel identification. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 1397-1400). Barcelona.

Assmann, P. F., & Nearey, T. M. (2003b). A review of results on frequency shifts and vowel identification. *Canadian Acoustics*, **31**, 18-19.

Assmann, P. F., & Nearey, T. M. (2007). Relationship between fundamental and formant frequencies in voice preference. *Journal of the Acoustical Society of America*, **122**, EL35-EL43. Retrieved May 8, 2009, from http://scitation.aip.org/journals/doc/JASMAN-ft/vol_122/iss_2/EL35_1.html. doi:10.1121/1.2719045

Assmann, P. F., Nearey, T. M., & Dembling, S. (2006). Effects of frequency shifts on perceived naturalness and gender information in speech. In *Proceedings of the 9th International Conference on Spoken Language Processing* (pp. 889-892). Pittsburgh.

Assmann, P. F., Nearey, T. M., & Scott, J. M. (2002). Modeling the perception of frequency-shifted vowels. In *Proceedings of the 7th International Conference on Spoken Language Processing* (pp. 425-428). Denver.

Bachorowski, J.-A., & Owren, M. J. (1999). Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *Journal of the Acoustical Society of America*, **106**, 1054-1063.

Bennett, S., & Montero-Diaz, L. (1982). Children's perception of speaker sex. *Journal of Phonetics*, **10**, 113-121.

Bladon, R. A. W. (1982). Arguments against formants in the auditory representation of speech. In R. Carlson & B. Granstrom (Eds.), *The representation of speech in the peripheral auditory system* (pp. 95-102). Amsterdam: Elsevier.

Bladon, R. A. W., & Lindblom, B. (1981). Modeling the judgment of vowel quality differences. *Journal of the Acoustical Society of America*, **69**, 1414-1422.

Childers, D. G., & Wu, K. (1991). Gender recognition from speech: Part II: Coarse analysis. *Journal of the Acoustical Society of America*, **90**, 1828-1840.

Chistovich, L. A., & Lublinskaja, V. V. (1979). The "center of gravity" effect in vowel spectra and critical distance between formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing Research*, **1**, 185-195.

Coleman, R. O. (1971). Male and female voice quality and its relationship to vowel formant frequencies. *Journal of Speech & Hearing Research*, **14**, 565-577.

Coleman, R. O. (1976). A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice. *Journal of Speech & Hearing Research*, **19**, 168-180. doi:10.1121/1.424675

de Cheveigné, A., & Kawahara, H. (1999). A missing-data model of vowel identification. *Journal of the Acoustical Society of America*, **105**, 3497-3508.

Disner, S. F. (1980). Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America*, **67**, 253-261.

Eklund, I., & Traunmüller, H. (1997). Comparative study of male and female whispered and phonated versions of the long vowels of Swedish. *Phonetica*, **54**, 1-21.

Fant, G. (1975). Non-uniform vowel normalization. *Speech Transmission Laboratories–Quarterly Progress & Status Report*, **2-3**, 1-19.

Fellowes, J. M., Remez, R. E., & Rubin, P. E. (1997). Perceiving the sex and identity of a talker without natural vocal timbre. *Perception & Psychophysics*, **59**, 839-849.

Fitch, J. L., & Holbrook, A. (1970). Modal vocal fundamental frequency of young adults. *Archives of Otolaryngology*, **92**, 379-382.

Fujisaki, H., & Kawashima, T. (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio & Electroacoustics*, **16**, 73-77.

Graddol, D., & Swann, J. (1983). Speaking fundamental frequency: Some physical and social correlates. *Language & Speech*, **26**, 351-366.

Henton, C. G. (1989). Fact and fiction in the description of female and male pitch. *Language & Communication*, **9**, 299-311.

Hillenbrand, J. M., & Gayvert, R. A. (1993). Vowel classification based on fundamental frequency and formant frequencies. *Journal of Speech & Hearing Research*, **36**, 694-700.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, **97**, 3099-3111.

Hillenbrand, J. M., & Houde, R. A. (2003). A narrow band pattern-matching model of vowel perception. *Journal of the Acoustical Society of America*, **113**, 1044-1055.

Hillenbrand, J. M., Houde, R. A., & Gayvert, R. A. (2006). Speech perception based on spectral peaks versus spectral shape. *Journal of the Acoustical Society of America*, **119**, 4041-4054.

Hillenbrand, J. M., & Nearey, T. M. (1999). Identification of resynthesized /hVd/ utterances: Effects of formant contour. *Journal of the Acoustical Society of America*, **105**, 3509-3523.

House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America*, **25**, 105-113.

Hudson, A. I., & Holbrook, A. (1981). A study of the reading fundamental vocal frequency of young black adults. *Journal of Speech & Hearing Research*, **24**, 197-200.

Ingemann, F. (1968). Identification of the speaker's sex from voiceless fricatives. *Journal of the Acoustical Society of America*, **44**, 1142-1144.

Ito, M., Tsuchida, J., & Yano, M. (2001). On the effectiveness of whole spectral shape for vowel perception. *Journal of the Acoustical Society of America*, **110**, 1141-1149.

Kawahara, H., Masuda-Kasuse, I., & de Cheveigné, A. (1999). Restructuring speech representations using pitch-adaptive time-frequency smoothing and instantaneous-frequency-based F_0 extraction: Possible role of repetitive structure in sounds. *Speech Communication*, **27**, 187-207.

Klatt, D. H. (1982). Speech processing strategies based on auditory models. In R. Carlson & B. Granstrom (Eds.), *The representation of speech in the peripheral auditory system* (pp. 181-196). Amsterdam: Elsevier.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, **87**, 820-857.

Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., & Bourne, V. T. (1976). Speaker sex identification from voiced, whispered, and filtered isolated vowels. *Journal of the Acoustical Society of America*, **59**, 675-678.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America*, **85**, 2114-2134.

Miller, R. L. (1953). Auditory tests with synthetic vowels. *Journal of the Acoustical Society of America*, **25**, 114-121.

Nearey, T. M. (1978). *Phonetic feature systems for vowels*. Bloomington: Indiana University Linguistics Club.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America,* **85**, 2088-2113.

Nearey, T. M., Hogan, J. T., & Rozsypal, A. J. (1979). Speech signals, cues, and features. In G. D. Prideaux (Ed.), *Perspectives in experimental linguistics* (pp. 73-96). Amsterdam: Benjamins.

Owren, M. J., Berkowitz, M., & Bachorowski, J.-A. (2007). Listeners judge talker sex more efficiently from male than from female vowels. *Perception & Psychophysics*, **69**, 930-941.

Paul, D. B. (1981). The spectral envelope estimation vocoder. *IEEE Transactions on Acoustics, Speech, & Signal Processing*, **29**, 786-794.

Peterson, G., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, **24**, 175-184.

Potter, R. K., & Steinberg, J. C. (1950). Toward the specification of speech. *Journal of the Acoustical Society of America*, **22**, 807-820.

Remez, R. E., & Rubin, P. E. (1984). On the perception of intonation from sinusoidal sentences. *Perception & Psychophysics*, **35**, 429-440.

Remez, R. E., & Rubin, P. E. (1993). On the intonation of sinusoidal sentences: Contour and pitch height. *Journal of the Acoustical Society of America*, **94**, 1983-1988.

Schwartz, M. F., & Rine, H. E. (1968). Identification of speaker sex from isolated, whispered vowels. *Journal of the Acoustical Society of America*, **44**, 1736-1737.

Slawson, A. W. (1968). Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency. *Journal of the Acoustical Society of America*, **43**, 87-101.

Smith, D. R. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *Journal of the Acoustical Society of America*, **118**, 3177-3186.

Smith, D. R. R., Walters, T. C., & Patterson, R. D. (2007). Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled, *Journal of the Acoustical Society of America*, **122**, 3628-3639.

Stoicheff, M. L. (1981). Speaking fundamental frequency characteristics of nonsmoking female adults. *Journal of Speech & Hearing Research*, **24**, 437-441.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, **79**, 1086-1100.

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *Journal of the Acoustical Society of America*, **85**, 1699-1707.

Traunmüller, H. (1984). Articulatory and perceptual factors controlling the age- and sex-conditioned variability in formant frequencies of vowels. *Speech Communication*, **3**, 49-61.

Traunmüller, H. (1988). Paralinguistic variation and invariance in the characteristic frequencies of vowels. *Phonetica*, **45**, 1-29.

Whiteside, S. P. (1998). The identification of a speaker's sex from synthesized vowels. *Perceptual & Motor Skills*, **87**, 595-600.

Zahorian, S. A., & Jagharghi, A. J. (1993). Spectral-shape features versus formants as acoustic correlates for vowels. *Journal of the Acoustical Society of America*, **94**, 1966-1982.

Zue, V. W., Seneff, S., & Glass, J. (1990). Speech database development at MIT: TIMIT and beyond. *Speech Communication*, **9**, 351-356.

NOTES

1. Sinusoidal sentences do, in fact, give the listener the impression of an intonation contour, albeit an odd one. Remez and Rubin (1984, 1993) showed that the intonation percept is related primarily to the frequency contour of the F_1 sinusoid. This contour, however, is entirely unrelated to the f_0 contour of the original utterance.

2. Fundamental frequency is also influenced by vowel identity, though in both absolute and relative terms the effect is much smaller than the influence of vowel identity on formant frequencies. On average, higher vowels are systematically higher in f_0 than lower vowels (e.g., House & Fairbanks, 1953).

3. The difficulty with the TIMIT sentences is not in measuring the formants, but rather in deriving the stable means and *SD*s for each vowel category that are used in the normalization procedure. These means and *SD*s are needed for the conversion of each individual token to a standard score that reflects primarily variations in vocal tract length and not phonetic context, speaking rate, sentential stress, lexical stress, or any number of other factors that, of necessity, vary freely in sentence material but not in /hVd/ syllables.

4. To cite just one example, Peterson and Barney (1952) reported 10,279 labeling responses to instances of /u/ spoken by the 76 talkers in their study, and there was not a single response of either /i/ or /ɪ/, errors that would be expected if closely spaced F_1 and F_2 merged to form a single peak. When /u/ was misheard, the response was nearly always /ʊ/.

(Manuscript received July 3, 2008;
revision accepted for publication March 7, 2009.)