

Time-varying spectral change in the vowels of children and adults

Peter F. Assmann and William F. Katz

*School of Human Development and Callier Center for Communication Disorders,
The University of Texas at Dallas, Box 830688, Richardson, Texas 75083*

(Received 30 August 1999; accepted for publication 23 June 2000)

Recent studies have shown that time-varying changes in formant pattern contribute to the phonetic specification of vowels. This variation could be especially important in children's vowels, because children have higher fundamental frequencies (f_0 's) than adults, and formant-frequency estimation is generally less reliable when f_0 is high. To investigate the contribution of time-varying changes in formant pattern to the identification of children's vowels, three experiments were carried out with natural and synthesized versions of 12 American English vowels spoken by children (ages 7, 5, and 3 years) as well as adult males and females. Experiment 1 showed that (i) vowels generated with a cascade formant synthesizer (with hand-tracked formants) were less accurately identified than natural versions; and (ii) vowels synthesized with steady-state formant frequencies were harder to identify than those which preserved the natural variation in formant pattern over time. The decline in intelligibility was similar across talker groups, and there was no evidence that formant movement plays a greater role in children's vowels compared to adults. Experiment 2 replicated these findings using a semi-automatic formant-tracking algorithm. Experiment 3 showed that the effects of formant movement were the same for vowels synthesized with noise excitation (as in whispered speech) and pulsed excitation (as in voiced speech), although, on average, the whispered vowels were less accurately identified than their voiced counterparts. Taken together, the results indicate that the cues provided by changes in the formant frequencies over time contribute materially to the intelligibility of vowels produced by children and adults, but these time-varying formant frequency cues do not interact with properties of the voicing source. © 2000 Acoustical Society of America.

[S0001-4966(00)01410-7]

PACS numbers: 43.71.An, 43.70.Ep, 43.71.Es [KRK]

I. INTRODUCTION

The perception of vowel quality is determined mainly by the formant pattern and its changes over time (Rosner and Pickering, 1994). Traditionally, vowels were described as static entities, analyzed in terms of a single, brief spectral sample taken from their central region or "nucleus" (e.g., Peterson and Barney, 1952). However, several sources of evidence now indicate that time-varying changes in the frequencies of the lowest three formants contribute to the perception of vowel quality, even in monophthongs (Di Benedetto, 1989; Fox, 1989; Hillenbrand and Gayvert, 1993; Hillenbrand *et al.*, 1995; Nábelek and Ovchinnikov, 1997; Nearey and Assmann, 1986; Nearey, 1989; Andruski and Nearey, 1992; Strange *et al.*, 1983; Strange, 1989; Pols and van Son, 1993; Zahorian and Jaghargi, 1991, 1993).

Hillenbrand (1995) and Hillenbrand and Nearey (1999) reported that vowels synthesized with "flattened" formant tracks (i.e., with the formant pattern held constant across the duration of the vowel) were identified less accurately than vowels for which the natural variations in formant frequencies were preserved. Their synthesized stimuli were modeled after a large sample of vowels produced by men, women, and children (ages 10–12 years). Formant flattening led to a 15% drop in mean identification accuracy, suggesting that formant movement plays an important role in the perceptual specification of American English vowels.

One reason for the detrimental effects of formant flat-

tening may be that formant movement helps to disambiguate pairs of vowels whose spectral shapes are similar in their "nucleus" regions but differ in their off-glides [e.g., the vowels /t/ and /e/ in American English (Nearey and Assmann, 1986)]. Formant-frequency changes could also provide evidence of the locations of "merged" formant peaks when pairs of formants approach one another in frequency. The likelihood of merged formants is greater when the fundamental frequency (f_0) is high and the spectrum envelope is sparsely "sampled" at the frequencies of the harmonics. Given the importance of the formant peaks for vowel identification, it might be predicted that vowels with high f_0 's would be identified less accurately than those with low f_0 's. This prediction has been confirmed in experiments with synthetic vowels (Ryalls and Lieberman, 1982; Diehl *et al.*, 1996).

An illustration of sparse spectral sampling is provided in Fig. 1, which shows the amplitude spectrum of the vowel /ɔ/ spoken by a 3-year old child. The left panel shows the amplitude spectrum of the onset portion of the vowel, along with the envelope of the spectrum estimated by linear predictive coding (LPC) analysis. The f_0 is about 262 Hz, and hence neither representation provides clear evidence of the second formant, located in the vicinity of 1050 Hz. This is a typical example of a pair of merged formants: only a single peak is evident in the region of $F1$ and $F2$.

In the right panel, the time variation in formant pattern is

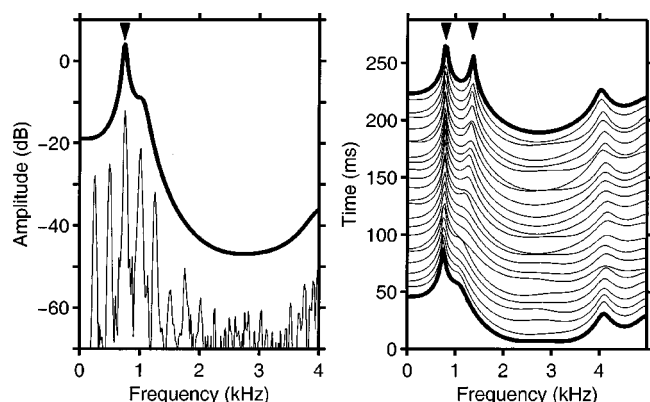


FIG. 1. Effects of time-varying changes in formant pattern on the resolution of formant peaks. The left panel shows the amplitude spectrum of a 45-ms segment from the onset portion of the vowel /ɔ/, spoken by a 3-year-old child. The dark trace is the spectrum envelope, estimated using autocorrelation LPC (10 coefficients, 512-point Hanning window, 10-kHz sample rate). The right panel shows a succession of LPC analyses obtained at 8-ms intervals throughout the vowel.

illustrated by a set of LPC spectra sampled at successive 8-ms intervals throughout the duration of the vowel. The envelope of the spectrum is more clearly defined, by extrapolation, when the vowel is sampled repeatedly at different time points. If formant movement is beneficial because it provides a basis for tracking merged formants, then these benefits should be greater for vowels with high f_0 's because they contain a higher incidence of such mergers.

The relevance of the example in Fig. 1 depends on the ability of the auditory system to recover features of the spectral envelope (such as the frequencies of formant peaks or aspect of spectral shape) from the raw waveform. Psycho-physical studies suggest that auditory frequency analysis is sufficiently selective at low frequencies to resolve the individual harmonic components of vowels in the frequency region below 1 kHz, even in adult male voices with low f_0 's (Moore and Glasberg, 1987). Since the first formant peak does not necessarily coincide with any single harmonic, its frequency must be derived, possibly by interpolating across the peaks in the excitation pattern that correspond to the individual harmonics (Darwin and Gardner, 1985; Assmann and Nearey, 1987). When the f_0 is high (as in children's speech), the spectrum envelope is less clearly defined in the speech signal, further increasing the uncertainty in the formant estimation process (Dissard and Darwin, 2000) and, by inference, the likelihood of making identification errors. If a higher f_0 leads to poorer specification of formant peaks or other relevant aspects of spectral shape, then time-varying changes in formant-frequency ought to provide greater benefits for children's vowels than for adults' vowels.¹

A corollary of this prediction is that time-varying changes in formant frequency should provide reduced benefits when f_0 is eliminated from the signal, as in whispered speech. Whispered speech retains the formant structure of voiced speech but not its harmonic fine structure. The broadband noise excitation of whispered speech generates a continuous spectrum, and potentially more accurate resolution of the formant peaks, provided the analysis window is sufficient in length to average out short-term fluctuations in the noise.²

If formant peaks are more accurately represented in the auditory excitation patterns of whispered than voiced vowels, then whispered vowels might be expected to show reduced benefits of formant movement, compared to voiced vowels. However, whispered vowels are less intelligible than voiced vowels under some conditions (Tartter, 1991; Katz and Assmann, 2000). Voicing source manipulations therefore provide a basis for testing the generality of the perceptual benefits of time-varying changes in formant pattern, and may yield further insights into the mechanisms involved in vowel perception.

The present study had two main objectives: first, to determine whether time-varying changes in formant frequencies make a greater contribution to the identification of children's vowels than adults' vowels; second, to determine how the perceptual effects of formant movement generalize across changes in voicing source. Listening tests were carried out using natural and synthesized vowels, including conditions where the center frequencies of formants were held constant over the duration of the vowel (Hillenbrand, 1995). Vowels were synthesized using pulsed excitation (generating voiced vowels) and noise excitation (generating whispered vowels).

II. EXPERIMENT 1

Compared to adults, children's speech is more variable in f_0 , formant frequencies, and durational properties (Eguchi and Hirsh, 1969; Kent, 1976; Smith *et al.*, 1995; Hillenbrand *et al.*, 1995; Lee *et al.*, 1999). Since the majority of earlier studies of children's speech have measured speech acoustics in older children, we included younger children of ages 7, 5, and 3 years in our sample.

A. Speech materials

Recordings were made of the 12 monophthongal vowels of American English in /hVd/ context: /i/ (heed); /ɪ/ (hid); /e/ (hayed); /ɛ/ (head); /æ/ (had); /ʌ/ (hud); /ɑ/ (hod); /ɔ/ (hawed); /ɜ/ (herd); /o/ (hoed); /ʊ/ (hood); /u/ (who'd).

B. Talkers

Ten men, 10 women, and 30 children (ages 7, 5, and 3) served as talkers. The majority of the adult talkers were long-time residents of the Dallas, Texas region, but 38% had lived in other cities during their childhood. All of the children were raised in the urban Dallas area.

C. Recording procedure

An audio cassette master tape with 6 repetitions of the 12 vowels in random sequence served as a prompt for the talkers. The adult talkers completed the set twice, for a total of 144 vowel tokens (12 vowels \times 12 repetitions of each vowel). Each recording session lasted about 40 min. Recording sessions for the children took somewhat longer, and hence children completed only one set of 72 vowel tokens (12 vowels \times 6 repetitions). Because children occasionally had difficulty producing target sounds following the first rep-

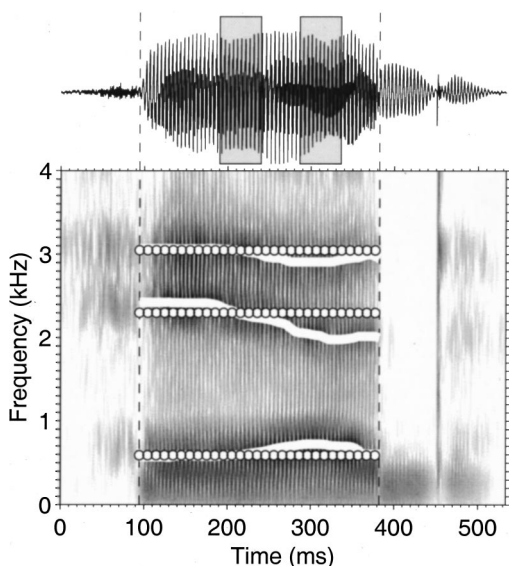


FIG. 2. The upper trace is the waveform of the syllable /hed/ spoken by an adult female. The shaded boxes indicate the first and second sample windows, starting at 33% and 66% of the distance from onset to offset of the vowel, respectively. The spectrogram shows the trajectories of F_1 , F_2 , and F_3 frequencies with white lines. The unfilled circles show the “flattened” formants used in the FlatF condition of experiment 1.

etition of the audio tape, they were given additional chances to repeat target items following examples given by the examiner.

Recordings were made in a sound-treated room using a Shure SM-94 microphone and a portable audio DAT recorder, model AIWA HD-X3000. The digital waveforms were transferred to computer disk at a rate of 48 kHz and 16-bit resolution using a DAT-Link+ digital audio interface.

D. Acoustic analysis

A subset of 180 vowel tokens (12 vowels \times 3 talkers \times 5 talker groups) was used to extract parameters for synthesizing the stimuli in experiment 1. Selection was based on informal judgments of adequate pronunciation quality by the two authors. The waveforms were digitally low-pass filtered with an eighth-order elliptical filter (5.7 kHz, -115 dB/oct) and resampled at 12 kHz for subsequent analyses. The onsets and offsets of the vowels were determined by visual inspection, as illustrated by the dashed lines in Fig. 2. Vowel onset was defined as the beginning of the first pitch period of the voiced segment of the syllable. Vowel offset was defined as the end of the last pitch period before the stop closure created by the final /d/.

To obtain a parametric description of the vowels suitable for a formant-based synthesizer, estimates of fundamental frequency (f_0), amplitude of voicing (AV) and formant center frequencies (F_1 - F_4) were obtained every 5 ms. Estimates of f_0 were made using an implementation of the Meddis and Hewitt (1991) pitch model. The amplitude of voicing (AV) was computed using a sliding rectangular root mean square (rms) window (25-ms frames, 5-ms overlap). Formant center frequencies were tracked using a custom MATLAB program (Assmann *et al.*, 1994). This program allows the user to identify the trajectories of the formants by visual inspection

and superimpose their tracks on a spectrogram of the vowel using a mouse-based drawing tool. A synthesized version can then be generated and played back to compare with the spoken version, and the process can be repeated until a good match is found.

E. Statistical analysis of vowel formant frequencies

To analyze formant-frequency change, we adopted a version of the “dual-target” model described by Nearey and Assmann (1986). Vowel formant frequencies were measured using LPC analysis and a semi-automatic formant-tracking procedure, described in detail in Sec. III A below. Measurements were taken at two different time points in the vowel, illustrated by the shaded boxes in Fig. 2. The sample windows had their onsets at 33% and 66% of the vowel’s duration, respectively. The goal was to include as much as possible of the formant movement within the vowel, while minimizing the neutralizing effects of the flanking consonants. Informal listening to gated versions of the vowels that spanned the two sample windows did not reveal a strong impression of the final /d/ for a majority of the tokens. Moreover, the formant trajectory defined by the two targets did not consistently point toward the same spectral locus (Delattre *et al.*, 1955; Sussman *et al.*, 1991) as might be expected if the second sample window were too close to the consonant closure. F_1 , F_2 , F_3 frequencies and f_0 were estimated as the median of five successive measurements spaced 5-ms apart.

Table I lists the f_0 and formant-frequency estimates from the first sample window for the 12 vowels and 5 talker groups. These measurements served as the basis for the “flattened formant” stimuli of experiment 1. Means and standard errors are based on three talkers in each group. An analysis of variance (ANOVA) of vowel formant frequencies was carried out, covering the factors vowel (with 12 levels), group (with 5 levels), and sample (with two levels: first and second sample windows). As expected, there were significant main effects of vowel for F_1 [$F_{(11,110)} = 84.59$, $p < 0.01$], F_2 [$F_{(11,110)} = 117.32$, $p < 0.01$], and F_3 [$F_{(11,110)} = 30.85$, $p < 0.01$]; and talker group for F_1 [$F_{(4,10)} = 12.18$, $p < 0.01$], F_2 [$F_{(4,10)} = 52.08$, $p < 0.05$], and F_3 [$F_{(4,10)} = 47.82$, $p < 0.01$], but the interaction of talker group \times vowel was not significant for any of the formants.

In addition to showing the expected pattern of higher formant frequencies in the vowels of young children, Table I shows that the standard errors are larger for several of the children’s vowels, compared to those of adults, consistent with reports of greater variability in children’s formant frequencies. However, it cannot be ruled out that this increased variability is due to formant measurement error associated with the higher f_0 ’s of children’s vowels, nor that this variability is irrelevant for vowel perception. These issues are addressed in experiments 1 and 2.

Of particular interest for the present study was the significant interaction of sample \times vowel, indicating the presence of time variation in the frequencies of F_1 [$F_{(11,110)} = 12.65$, $p < 0.01$] and F_2 [$F_{(11,110)} = 21.74$, $p < 0.01$], but not F_3 [$F_{(11,110)} = 1.14$, $p = 0.34$]. Figure 3 illustrates the pattern of formant movement for the male talkers. The mean F_1 and F_2 frequencies for the first and second time samples are

TABLE I. Means and standard errors (in parentheses) of vowel fundamental and formant frequencies (in Hz) from the first sample window, across the three talkers in each group. (See text for details.)

| | | /i/ | /u/ | /e/ | /ɛ/ | /æ/ | /ɒ/ | /ɜ-/ | /a/ | /ɔ/ | /o/ | /ʊ/ | /u/ |
|---------------|-------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| Adult Males | F3 | 3003 (61) | 2654 (64) | 2557 (39) | 2643 (69) | 2580 (30) | 2539 (99) | 1686 (41) | 2468 (53) | 2564 (111) | 2390 (86) | 2364 (83) | 2321 (112) |
| | F2 | 2345 (146) | 1974 (86) | 1982 (74) | 1855 (96) | 1809 (26) | 1455 (14) | 1457 (71) | 1214 (12) | 1081 (17) | 1182 (28) | 1376 (105) | 1373 (139) |
| | F1 | 300 (6) | 445 (31) | 497 (14) | 534 (21) | 694 (50) | 638 (49) | 523 (29) | 754 (49) | 654 (15) | 523 (31) | 426 (20) | 353 (20) |
| | f_0 | 110 (9) | 108 (12) | 111 (13) | 102 (10) | 101 (8) | 102 (10) | 105 (11) | 103 (8) | 101 (10) | 112 (10) | 112 (10) | 131 (7) |
| Adult Females | F3 | 3256 (194) | 2965 (177) | 2990 (210) | 2929 (158) | 2875 (97) | 2887 (153) | 1870 (84) | 2966 (16) | 2947 (125) | 2634 (129) | 2734 (94) | 2636 (136) |
| | F2 | 2588 (164) | 2161 (142) | 2309 (126) | 2144 (104) | 2051 (90) | 1751 (55) | 1508 (62) | 1273 (92) | 1203 (48) | 1470 (69) | 1685 (77) | 1755 (157) |
| | F1 | 429 (26) | 522 (58) | 572 (20) | 586 (42) | 836 (50) | 767 (19) | 640 (61) | 688 (32) | 816 (44) | 636 (24) | 516 (51) | 430 (28) |
| | f_0 | 216 (6) | 207 (8) | 209 (10) | 204 (10) | 199 (4) | 199 (7) | 201 (1) | 208 (9) | 194 (6) | 201 (8) | 207 (7) | 217 (12) |
| Age 7 | F3 | 3977 (161) | 3896 (226) | 3620 (42) | 3713 (244) | 3621 (318) | 3443 (160) | 2297 (97) | 3083 (189) | 3343 (178) | 3252 (87) | 3663 (177) | 3162 (303) |
| | F2 | 3402 (95) | 2825 (106) | 2822 (48) | 2485 (67) | 2324 (52) | 1896 (72) | 1776 (80) | 1565 (90) | 1494 (74) | 1601 (45) | 2031 (191) | 1838 (80) |
| | F1 | 358 (46) | 583 (63) | 590 (31) | 799 (61) | 1074 (89) | 832 (140) | 601 (41) | 954 (128) | 895 (54) | 620 (57) | 579 (71) | 491 (25) |
| | f_0 | 257 (8) | 237 (6) | 253 (12) | 246 (6) | 235 (16) | 250 (14) | 253 (6) | 254 (4) | 241 (3) | 246 (8) | 250 (8) | 257 (6) |
| Age 5 | F3 | 4058 (83) | 3954 (95) | 3923 (56) | 3922 (106) | 4022 (130) | 3742 (159) | 2498 (131) | 3136 (246) | 3228 (249) | 3133 (312) | 3626 (167) | 3809 (90) |
| | F2 | 3535 (74) | 2914 (139) | 3050 (61) | 2684 (72) | 2505 (52) | 1965 (156) | 2019 (53) | 1602 (68) | 1502 (58) | 1620 (110) | 1919 (192) | 1711 (172) |
| | F1 | 472 (81) | 571 (47) | 580 (25) | 871 (102) | 1161 (52) | 732 (77) | 599 (52) | 1066 (87) | 850 (144) | 647 (20) | 526 (20) | 471 (69) |
| | f_0 | 280 (13) | 269 (11) | 269 (8) | 263 (8) | 240 (3) | 258 (6) | 251 (13) | 251 (7) | 241 (3) | 254 (4) | 270 (11) | 273 (15) |
| Age 3 | F3 | 4061 (188) | 4331 (177) | 3721 (277) | 4294 (145) | 3961 (440) | 3333 (313) | 2661 (147) | 3866 (427) | 3382 (385) | 3228 (276) | 3899 (282) | 2866 (314) |
| | F2 | 3437 (67) | 2740 (62) | 2863 (72) | 2639 (13) | 2503 (170) | 1965 (58) | 1752 (94) | 1656 (109) | 1465 (202) | 1636 (84) | 1828 (205) | 1891 (216) |
| | F1 | 427 (42) | 621 (44) | 717 (37) | 760 (10) | 1256 (110) | 789 (69) | 726 (60) | 1060 (109) | 938 (111) | 652 (33) | 649 (10) | 502 (45) |
| | f_0 | 246 (29) | 230 (29) | 218 (27) | 211 (16) | 227 (14) | 214 (14) | 233 (9) | 248 (21) | 209 (26) | 227 (36) | 251 (21) | 271 (21) |

shown by the origins and tails of the arrows. The direction and extent of formant-frequency movement is consistent with studies of other dialects of American English (e.g., Nearey and Assmann, 1986, for western Canada; Hillenbrand and Nearey, 1999, for western Michigan). The greatest formant movement is seen for the vowels /e/, /o/, /a/, /ɒ/, /u/, and /ɔ/, while /i/, /u/, /æ/, and /ɜ-/ are relatively stationary. In general, the lax vowels /ɪ/, /ɛ/, and /æ/ tend to point toward the middle of the vowel space, consistent with the pattern found in other dialects of American English, although the magnitude of these changes is smaller for the north Texas vowels. This discrepancy may reflect dialect variation, or differences in the choice of time windows. Our first sample point was taken somewhat later, and the second sample somewhat earlier than in previous studies.³

The pattern of formant movement was similar across the five talker groups: neither the talker group \times sample nor the talker group \times sample \times vowel interactions were significant for any of the formants. Hence children's vowels appear to

display qualitatively similar patterns of vowel-inherent spectral change to those of adults.

F. Experimental conditions

The stimuli were natural and synthesized versions of the 12 vowels provided by the adult and child talkers. Stimuli were presented to adult listeners for identification in five conditions:

- (1) *Natural*: The vowel portion of the syllable was extracted (as illustrated by the dashed lines in Fig. 2) and presented without further modification.
- (2) *Full*: A synthesized version of the vowel was constructed using the acoustic parameters (f_0 , AV, F1, F2, F3, F4, F5) derived from an analysis of the vowel portion of the /hVd/ syllables. Synthetic vowels were constructed using a frequency-domain version of the Klatt (1980) cascade formant synthesizer (Assmann *et al.*, 1994) with six formants at a sample rate of 24 kHz. The

center frequencies of the higher formants were fixed at 4500 Hz (F_5) and 5500 Hz (F_6) for the adult males, using the synthesizer defaults for an adult male voice suggested by Klatt (1980). The frequencies of F_5 and F_6 were scaled upwards by 20% for the adult females (5400 and 6600 Hz, respectively) and by 50% for the three groups of children (6750 and 8250 Hz). The bandwidths of the six formants were held constant (B_1 : 90; B_2 : 110; B_3 : 170; B_4 : 250; B_5 : 300; B_6 : 450 Hz). Stimuli were scaled to the maximum peak-to-peak amplitude afforded by the 16-bit quantization range. A 10.7-ms Kaiser window was used to shape the onset and offset of each stimulus to minimize transients.

- (3) *FlatF1*: The center frequency of the first formant was held constant at the estimated F_1 frequency from the first sample window (indicated by the first shaded box in Fig. 2). The remaining parameters displayed “natural” (i.e., measured) variations as a function of time.
- (4) *FlatF2*: The center frequency of the second formant was held constant at the estimated F_2 frequency from the first sample window.
- (5) *FlatF*: The center frequencies of all formants and f_0 were held constant at their estimated frequencies from the first sample window.

G. Listeners

The listeners were nine adults who reported normal hearing and were native speakers of American English. All had completed an undergraduate course in phonetics. Most were long-term residents of the Dallas, Texas area.

H. Procedure

The stimuli were presented on-line at a sample rate of 24 kHz, low-pass filtered (10-kHz cutoff; Tucker-Davis Technologies FT5), attenuated (TDT PA4), amplified (TDT HB5), and presented diotically via headphones (Sennheiser HD-414) in a double-walled sound booth. Stimulus presentation levels ranged from 69 to 81 dB SPL (A weighting), with a mean of 74 dB. Stimuli representing the 12 vowels, 5 talker groups, and 5 synthesis conditions were interspersed in random order from trial to trial. Listeners were tested individually, and the experiment was self-paced. They responded to each stimulus by selecting 1 of 12 panels from a response box displayed on the computer screen. The panels were labeled with the orthographic representations of the /hVd/ words, and their corresponding phonetic symbols were superimposed on the screen with a plastic overlay.

Listeners began by completing three brief practice sets. In each set they heard vowels spoken by different talkers, one exemplifying each talker group, but using talkers who were not included in the main experiment. In the first set they listened to examples of 12 natural vowels without assigning vowel responses. In the second set they identified 36 vowels from the natural condition and received feedback indicating whether their responses were correct or incorrect. If their score was less than 85% correct (31/36) they were required to repeat the set until they reached this goal. In the third set they heard examples of 12 synthesized vowels from

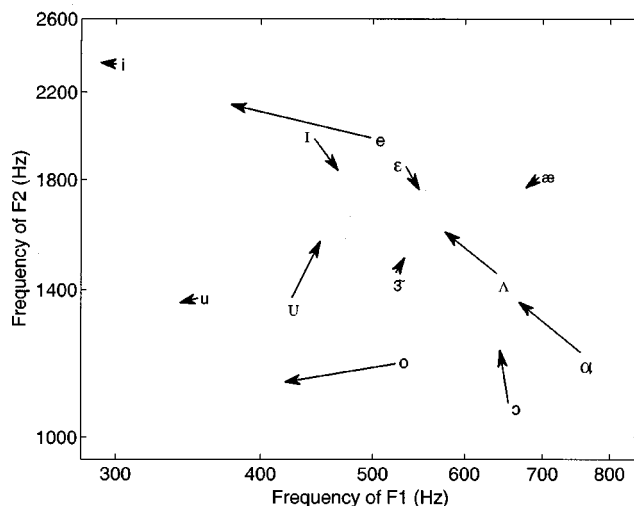


FIG. 3. Formant-frequency movement for the 12 vowels spoken by the adult males. The origin of each arrow shows the mean F_1 and F_2 frequencies across the three talkers for the first sample window (starting at 33% of the vowel duration), while the tail shows the second sample window (starting at 66% of the vowel duration).

the full condition without assigning vowel responses. They next completed the main experiment, which included 1260 stimuli [12 vowels \times 5 groups \times 3 talkers per group \times 7 conditions, i.e., the five conditions described above plus two additional conditions described elsewhere (Katz and Assmann, 2000)] in three blocks of 420 trials. Each block lasted about 40 min.

I. Results and discussion

The identification data were analyzed with two main objectives: (i) to compare identification accuracy for natural and synthesized vowels spoken by adults and children, and (ii) to study the effects of “flattening” the formant trajectories on vowel identification for the five talker groups. A three-way analysis of variance (ANOVA) revealed significant main effects of talker group [$F_{(4,32)} = 16.60$; $p < 0.01$], condition [$F_{(4,32)} = 138.00$, $p < 0.01$], and vowel [$F_{(11,88)} = 19.10$, $p < 0.01$]. All interactions were significant: talker group \times condition [$F_{(16,128)} = 5.01$; $p < 0.01$], condition \times vowel [$F_{(44,352)} = 17.50$; $p < 0.01$], talker group \times vowel [$F_{(44,352)} = 17.50$; $p < 0.01$], and talker group \times condition \times vowel [$F_{(176,1408)} = 3.38$; $p < 0.01$]. The key patterns in these data are described below.

1. Identification of natural and synthesized vowels

Identification accuracy was lower for the synthesized vowels (full condition) than for the spoken vowels (natural condition) after which they were patterned. Figure 4 shows this pattern across the five talker groups. The largest decline in accuracy was found for adult females (20%). The smallest decline was for the 3-year-old children (3%), whose natural recorded vowels were the least intelligible overall. The three-way interaction of talker group \times condition \times vowel was analyzed with Scheffé tests to assess the drop in intelligibility from natural to full conditions. The analysis revealed significantly ($p < 0.05$) lower accuracy in the full condition for

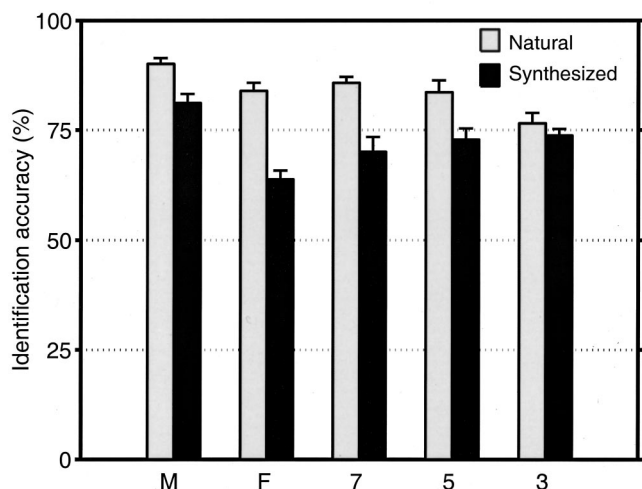


FIG. 4. Mean identification accuracy for the natural and full (synthesized) conditions for vowels spoken by adult males (M), adult females (F), and children (ages 7, 5, 3) in experiment 1. Error bars show the standard errors across the nine listeners.

the following talker group and vowel combinations: 3-year-olds: /o/; 5-year-olds: /e/, /u/; 7-year-olds: /u/; adult females: /ɪ/, /æ/, /ʊ/; adult males: /ɜ/. In some of these cases, the lower intelligibility of the synthesized vowels can be attributed to difficulties in tracking closely spaced formants.

2. Effects of holding formant frequencies constant

Significantly lower identification accuracy was obtained when either $F1$, $F2$, or all formants plus f_0 were held constant. Figure 5 shows mean accuracy in the four synthesis conditions for the five talker groups. On average, eliminating the time variation in $F1$ resulted in a 6% drop in identification. Holding $F2$ constant lowered performance by about 5%, while removing the time variation from all formant frequencies plus f_0 lowered performance by about 12%. These patterns were broadly similar across talker groups, with two

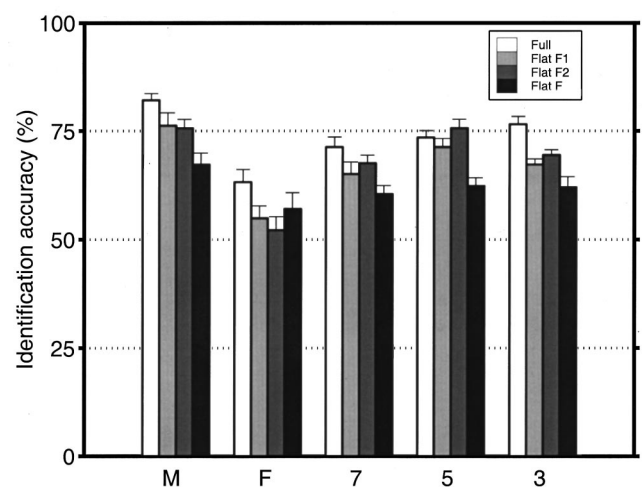


FIG. 5. Mean identification accuracy for vowels synthesized with time-varying formants (full condition) and with flattened formants (flat $F1$, flat $F2$, and FlatF conditions) spoken by adult males (M), adult females (F), and children (ages 7, 5, 3) in experiment 1. Error bars show the standard errors across the listeners.

exceptions. First, the 5-year-olds did not show a decline when $F1$ or $F2$ were held constant, although their scores were close to those of other talker groups when all formants and f_0 were held constant. Second, the adult females did not show a further decline when all formants and f_0 were held constant, compared to $F1$ and $F2$ alone. These differences merit further study with a larger sample of talkers. Overall, however, the results provided no support for the idea that time-varying changes in formant frequencies offer greater benefits for identifying vowels spoken by children.

Figure 6 shows that the effects of formant flattening were not evenly distributed across the vowels. The vowels /e/ and /o/, which are strongly diphthongized in American English, showed substantial declines, with more than a 65% drop from the full condition to FlatF condition. For both /e/ and /o/, holding $F1$ constant had a larger effect than holding $F2$ constant. Smaller declines were observed for /u/ and /ɔ/, while most of the remaining vowels showed little effect of formant flattening. Surprisingly, /a/ showed an *increase* of more than 20% when all formants and f_0 were held constant. The reason may be related to dialect and/or formant-tracking errors: for this vowel, eliminating the time variation made it easier to identify.

The results in Fig. 6 are combined across the five talker groups, but the ANOVA also showed a significant three-way interaction of talker group \times condition \times vowel. Interpretation of the three-way interaction was complicated by performance near ceiling or floor in several conditions, but overall there were only minor deviations from the pattern in Fig. 6. In general, deviations from the pattern were least evident for those vowels showing the greatest effects of formant flattening (i.e., /e/ and /o/).

3. Relationship between f_0 and formant flattening

The results of experiment 1 failed to confirm the prediction that time-varying changes in formant frequencies would provide greater benefits for the vowels of children than adults. One interpretation is that these benefits are not related to spectral resolution, but arise for other reasons. Before accepting this conclusion, however, it is necessary to consider other differences between the vowels of children and adults that might contribute to formant resolution. For example, narrowing the bandwidth of a formant or increasing its distance from adjacent formants could lead to improved specification of its peak, counteracting the loss of resolution caused by the high f_0 .

Formant bandwidths were held constant across the talker groups, ruling this out as a possible confound. Moreover, formant bandwidth estimates from natural speech generally suggest wider rather than narrower bandwidths with increasing formant-frequency above 500 Hz (Hawks and Miller, 1995), and formant bandwidths are estimated to be about 25% larger in women than in men (Fujimura and Lindqvist, 1971). A second factor is that children have smaller vocal tracts than adults, and hence the upward shift in their formant frequencies could lead to larger frequency distances between adjacent formants. However, a statistical analysis of formant distances ($F3-F2$ and $F2-F1$, with formant frequencies ex-

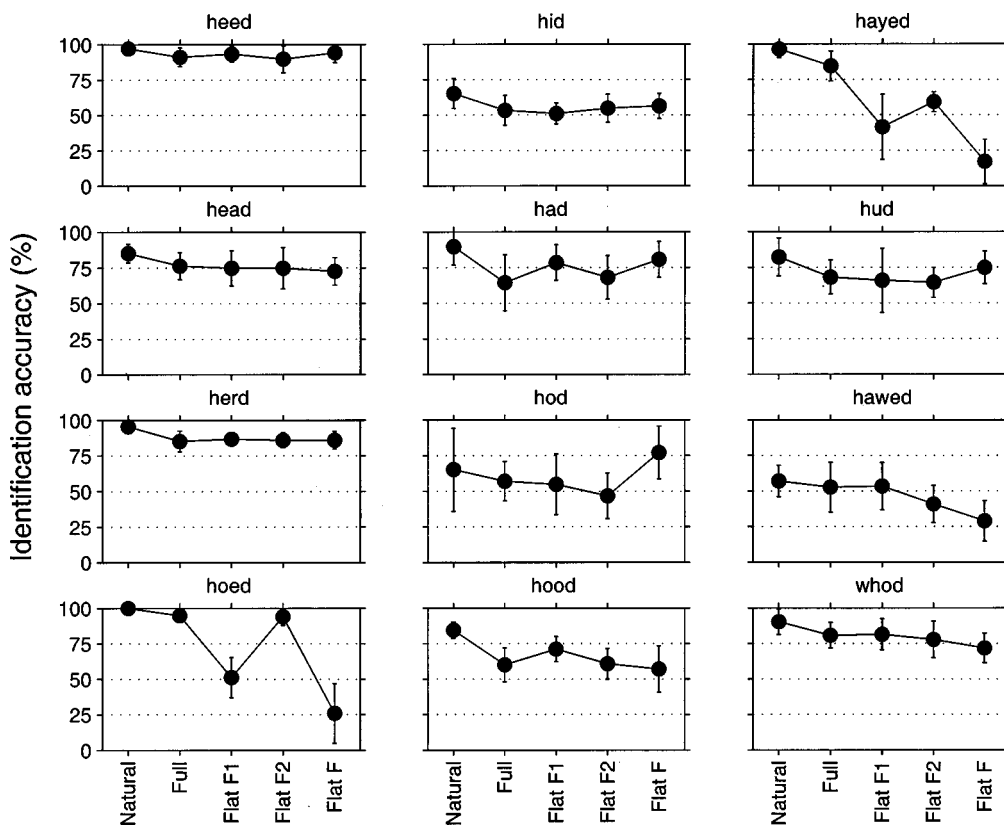


FIG. 6. Mean identification accuracy for the 12 natural vowels and synthesized versions, with time-varying formants (full condition) and with flattened formants (flat F1, flat F2, and FlatF conditions) in experiment 1. Error bars show the standard errors across the listeners.

pressed in log units), failed to reveal significant differences between child and adult talkers.

Although other differences between children's and adults' vowels cannot be ruled out, the results of experiment 1 suggest that the benefits of formant flattening are not enhanced in vowels with high f_0 's. This conclusion is supported by an analysis of individual vowel tokens. The amount of benefit provided by spectral change was quantified in terms of the difference between scores in the full condition and in the FlatF condition. The correlation between difference scores and the vowel's f_0 was close to zero ($r = -0.06$; $N=180$). There was a weak relationship between f_0 and formant distance, for $F2-F1$ ($r=0.29$; $N=180$; $p<0.01$) and $F3-F2$ ($r=0.26$; $N=180$; $p<0.01$). A partial correlation analysis was performed to assess the relationship between f_0 and formant flattening when the effects of formant separation are partialled out. In both cases the correlation remained near zero, for $F2-F1$ ($r=0.05$; $N=180$) and $F3-F2$ ($r=-0.01$; $N=180$). These analyses support the conclusion that the effects of formant flattening do not depend on f_0 .

III. EXPERIMENT 2

Because formants are relatively difficult to estimate in speech produced with high f_0 values, it is possible that some of the group-specific differences noted in experiment 1 (e.g., the large difference between natural vs full conditions for adult females and the small difference between natural versus full conditions for 3-year-old children) could have re-

sulted, in part, from errors in estimating formant frequencies by hand. Experiment 2 therefore evaluated a synthesis strategy that relied on a semi-automatic formant-tracking procedure.

A. Method

Experimental conditions were identical to those used in experiment 1, except that the stimuli were generated from the same recorded /hVd/ words using a semi-automated formant-tracking procedure. The frequencies of the formants ($F1$, $F2$, $F3$) were estimated using a dynamic programming method which assigns raw formant-frequency "candidates" to "tracks" given a set of constraints on the proximity of adjacent formants and continuity over time. Formant candidates were obtained by autocorrelation LPC (12-kHz sample rate, 512-point Hanning window, 5-ms frame update rate, and 95% pre-emphasis). The order of the LPC analysis was adjusted to match the number of expected formants in the 0-6 kHz range, and the number of coefficients ranged from 8 to 14, depending on the talker. Formant candidates were obtained by solving for the roots of the predictor polynomial. A subsequent stage of post-processing was included to eliminate residual errors due to "spurious" formant-frequency candidates. Finally, median smoothing was applied to eliminate sudden changes in formant-frequency.

Stimulus presentation, testing procedures, response collection, and data analysis were the same as described for experiment 1.

B. Listeners

The listeners were ten adults who reported normal hearing and were native speakers of American English. Most of the listeners were long term residents of the Dallas, Texas area. All had completed an undergraduate course in phonetics. One of the listeners had participated in experiment 1.

C. Results

There was little effect of formant-frequency estimation method: on average, hand-tracked and auto-tracked vowels showed less than 1% difference in intelligibility. The largest improvement was a nearly 6% difference in the female auto-tracked vowels. The largest drop was a 5% difference for 3-year-olds. A three-way ANOVA revealed the same pattern as in experiment 1, with a significant interaction of talker group \times condition \times vowel [$F_{(176,1584)}=3.78$; $p<0.01$]. Qualitative inspection of individual vowel \times talker group plots revealed an interaction similar to that found for experiment 1. Therefore, it appears that the perceptual consequences of the formant manipulations presented in experiment 1 do not depend on the formant-tracking method.

IV. EXPERIMENT 3

Experiments 1 and 2 showed that eliminating the time variation in formant frequencies from synthesized vowels led to a significant decline in intelligibility. However, no evidence was found of a link between the benefits of formant dynamics and f_0 . A significant correlation would be predicted if formant movement helps to define poorly resolved formant peaks or other aspects of spectral shape when f_0 is high.

Research with natural speech has shown that whispered vowels are less intelligible than voiced vowels (Tartter, 1991). Tartter found that the increased error rate for whispered vowels was due in part to increased confusions among vowels adjacent in the vowel space defined by $F1$ and $F2$ frequencies. Formant movement can help to disambiguate such pairs of vowels, and hence formant flattening might be expected to lead to a greater reduction in identification accuracy for whispered compared to phonated vowels. On the other hand, if formant movement is beneficial because it helps to define formant peaks when f_0 is high, then replacing the voicing source with noise might be expected to yield reduced effects of formant flattening in whispered vowels compared to voiced vowels, particularly for children. Experiment 3 tested these two contrasting predictions by examining the interaction of formant flattening and the presence or absence of voicing.

A. Method

Vowels were synthesized using the auto-tracked formant measurements from experiment 2 and the cascade formant synthesis model (Klatt, 1980). Vowels were synthesized ei-

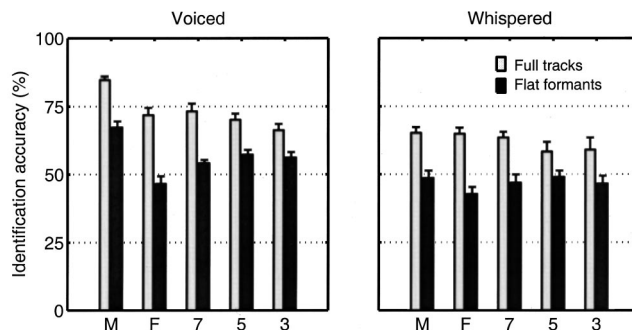


FIG. 7. Mean identification accuracy for the synthesized voiced and whispered vowels spoken by adult males (M), adult females (F), and three groups of children (ages 7, 5, 3) in experiment 3. Error bars show the standard errors across the eight listeners.

ther with pulsed excitation (to produce “voiced” vowels, as in experiment 2) or noise excitation (generating “unvoiced” or whispered vowels). In addition, each was synthesized with time-varying formant frequencies (full condition) or with all formants and f_0 held constant across time (FlatF condition). A total of 720 vowels was generated (2 voicing states \times 2 formant change conditions \times 5 talker groups \times 3 talkers/group \times 12 vowels). Stimulus presentation, testing procedures, response collection, and data analysis were the same as described for experiment 1.

B. Listeners

Eight adults who reported normal hearing and were native speakers of American English served as listeners. Most were long-term residents of the Dallas, Texas area. All had completed an undergraduate course in phonetics. Two of the listeners had participated in experiment 1, and one in experiment 2.

C. Results

Elimination of time variation in the frequencies of the formants and f_0 lowered identification accuracy by 17% for the voiced vowels and 15% for the unvoiced (whispered) vowels, on average. This shows that time-varying changes in formant frequencies have similar effects for whispered and voiced vowels. In comparison, replacement of the voicing source with aperiodic noise lowered mean identification accuracy by 11%.

There was a significant interaction of formant flattening \times talker group [$F_{(4,28)}=5.35$; $p<0.01$], and of voicing source \times talker group [$F_{(4,28)}=4.73$; $p<0.01$], but the three-way interaction of formant flattening \times voicing source \times talker group was not significant. Figure 7 shows that the shift from voicing to whisper has the effect of eliminating the selective advantage of the vowels of adult males over the other four talker groups.

Holding the formant frequencies constant had a larger effect on the vowels of adult females (24% drop) compared to those of other talker groups (11%–18% drop), possibly because the vowels of the adult females were generally more susceptible to synthesis errors (Fig. 4). Figure 7 shows that

TABLE II. Means and standard errors (in parentheses) of vowel identification accuracy for the eight listeners in experiment 3. The data are pooled across the five talker groups. Full tracks refers to vowels synthesized with time-varying formant frequencies and f_0 ; flat formants refers to vowels synthesized with steady-state formant frequencies and f_0 . Voiced vowels are synthesized with a pulsed voicing source; whispered vowels have broad-band noise excitation.

| | Full tracks | | Flat formants | |
|------|-------------|------------|---------------|------------|
| | Voiced | Whispered | Voiced | Whispered |
| /i/ | 84.2 (5.2) | 72.5 (4.8) | 83.3 (4.9) | 78.3 (4.7) |
| /ɪ/ | 74.2 (4.4) | 68.3 (4.8) | 40.0 (6.5) | 60.8 (6.6) |
| /e/ | 95.0 (2.1) | 75.8 (6.8) | 25.0 (6.0) | 3.3 (3.3) |
| /ɛ/ | 74.2 (3.7) | 66.7 (4.7) | 75.0 (4.8) | 65.0 (3.5) |
| /æ/ | 73.3 (4.9) | 37.5 (8.0) | 71.7 (4.5) | 44.2 (6.7) |
| /ʌ/ | 40.8 (5.7) | 38.3 (5.3) | 46.7 (5.2) | 35.0 (4.1) |
| /ɜ:/ | 94.2 (2.0) | 84.2 (3.5) | 85.0 (3.5) | 69.2 (3.8) |
| /ɑ/ | 43.3 (6.8) | 38.3 (7.5) | 51.7 (5.7) | 46.7 (4.9) |
| /ɔ/ | 55.8 (8.0) | 45.0 (5.9) | 38.3 (5.2) | 38.3 (5.5) |
| /o/ | 94.2 (2.0) | 85.8 (4.8) | 31.7 (6.6) | 11.7 (3.5) |
| /u/ | 72.5 (6.0) | 78.3 (5.0) | 44.2 (8.2) | 56.7 (7.5) |
| /ʊ/ | 77.5 (4.2) | 55.8 (5.5) | 83.3 (2.8) | 51.7 (4.3) |
| Mean | 73.3 (5.2) | 62.2 (4.8) | 56.3 (4.9) | 46.7 (4.7) |

eliminating the time variation from the formant frequencies led to a substantial drop in accuracy for all talker groups, with similar reductions for whispered and voiced vowels.

The three-way interaction of formant flattening \times voicing source \times vowel was significant [$F_{(11,77)} = 2.49$; $p < 0.01$] and is shown in Table II. Consistent with the results of experiment 1, the largest effects of formant flattening were observed for the vowels /e/ and /o/. Effects of formant flattening were similar for the voiced and whispered versions of these vowels. Several other vowels (e.g., /ɪ/ and /ʊ/) also showed small declines when their formants were held constant.

Although most vowels showed some effect of shifting from a voiced to whispered source, the changes were not uniform across the vowels. The largest effects emerged for the vowels /u/ and /æ/, neither of which was affected substantially by the flattening of their formants. Some vowels, like /i/ and /ɛ/, showed little effect of either manipulation.

Overall, the results of experiment 3 showed that the formant flattening and voicing source manipulations did not interact. Moreover, their effects were not the same for different vowels and talker groups. The ANOVA showed significant interactions of formant flattening \times talker group \times vowel [$F_{(44,308)} = 3.36$; $p < 0.01$] and voicing source \times talker group \times vowel [$F_{(44,308)} = 2.95$; $p < 0.01$], but no systematic patterns emerged from an inspection of these interactions. The four-way interaction of formant flattening \times voicing source \times talker group \times vowel was not significant. An interesting finding was that the vowels most affected by the change in voicing were hardly affected by the flattening of their formants. The reverse pattern did not appear to hold, since the vowels most affected by flattening their formants also showed a decline when the source changed from voicing to whisper.

The lower intelligibility of noise-excited (whispered) compared to pulse-excited (voiced) vowels was also found by Katz and Assmann (2000) using an experimental design

similar to experiment 3. They showed that this effect does not arise simply because whispered speech contains less energy at low frequencies, compared to voiced speech (Stevens, 1999): a similar reduction in intelligibility was found with noise-excited vowels which preserved the spectral tilt characteristics of the voiced vowels. Katz and Assmann attributed the reduced intelligibility of whispered vowels to the elimination of harmonicity and/or periodicity.

V. GENERAL DISCUSSION

An acoustic analysis confirmed previous findings (Egushi and Hirsh, 1969; Lee *et al.*, 1999) that children's vowels have higher formant frequencies than those of adults, as well as greater variability across talkers. The increased variability can be attributed to limitations in motor control and the reduced flexibility that accompanies the early stages of neuromotor maturation. In contrast, comparisons of formant-frequency trajectories for individual vowels revealed similar patterns for children and adults. These data suggest that children can produce appropriate patterns of formant-frequency movement in a mature manner as early as age 3.⁴

Three important findings emerged from the perceptual experiments. First, we confirmed that the removal of time-varying changes in formant frequency from synthetic vowels produced lower identification accuracy. This finding replicated earlier work by Hillenbrand (1995) and Hillenbrand and Nearey (1999) for W. Michigan talkers. Our talkers were from the Dallas, Texas region, extending the results to another dialect of American English, and also to children as young as age 3.

Eliminating the time variation from all formants and f_0 resulted in a 12%–16% reduction in mean identification accuracy in experiments 1–3. In comparison, flattening just one formant ($F1$ or $F2$) led to a drop of nearly half that magnitude (6% and 5% reduction, respectively). The similar drop for $F1$ and $F2$ suggests that time-varying changes in each formant made roughly the same contribution to vowel identification. However, differences were noted for /e/ and /o/, where flattening $F1$ degraded performance more than flattening $F2$. This pattern is consistent with the extent of acoustic change in the $F1 \times F2$ space, and is compatible with the finding that $F1$ frequency is more discriminable than $F2$ (Kewley-Port and Watson, 1994). $F2$ flattening had little effect on /o/, but rather pronounced effects on /e/. The smaller contribution of $F2$ for the vowel /o/ might be part of its phonetic specification in the north Texas dialect of American English. The proximity of $F1$ and $F2$ frequencies for /o/ might also be a factor.

The second finding was that the synthesized vocalic portions of the syllables were less well identified than the natural versions, despite the fact that the formant transitions and f_0 variation were preserved. This finding, also reported by Hillenbrand and Nearey (1999), suggests that although formant-frequency variation is indisputably important for vowel quality, other signal properties that contribute to vowel identification are lost when vowels are synthesized in accordance with the cascade formant synthesis model described by Klatt (1980). We are currently investigating

sources of error in the synthesis procedure that might be responsible for differences between the natural and synthesized vowels. Candidates include inaccuracies in the voicing source model, errors in formant measurement, and distortions of spectral shape resulting from the use of the formant representation (Hillenbrand and Nearey, 1999).

The third finding is that formant-frequency flattening did not interact with manipulations of the voicing source. Its effects were similar for adults and children, and no evidence was found that time-varying changes in formant-frequency provided greater benefits for children's voices with high f_0 . Flattening the trajectories of the formants led to similar reductions in identification accuracy for whispered and voiced vowels. Taken together, these results argue against the idea that time-varying changes in formant-frequency contribute by specifying the locations of formant peaks when f_0 is high.

Katz and Assmann (2000) tested a related hypothesis, that time variation in f_0 might help to trace out the spectrum envelope in vowels with high f_0 . Hillenbrand and Gayvert (1993) had previously shown a small but reliable improvement when steady-state vowels were synthesized with a linear f_0 glide rather than constant f_0 . However, using the same design as in experiment 1 of the current study, Katz and Assmann found no effect of replacing the natural time-varying changes in f_0 with steady-state f_0 . Moreover, there was little evidence of a link between vowel identification accuracy and f_0 , a result also reported by Hillenbrand and Nearey (1999). One explanation for the discrepancy between studies may be that time-varying changes in f_0 lead to higher identification accuracy only in the special case where vowels are synthesized with steady-state formant patterns, a condition not investigated in the present study. On the whole, these studies provided little evidence that time-varying changes in either f_0 or formant-frequency lead to improved resolution of spectral features such as formant peaks.

The finding that the benefits of time-varying changes in formant frequency do not interact with properties of the voicing source raises the following question: if formant movement does not help to specify the locations of formant peaks, how then do listeners overcome the problem of reduced spectral resolution when f_0 is high? The answer may be that the auditory representations of vowels with high f_0 are *not* subject to the degradation observed in their amplitude spectra. Accurate vowel identification may not require precise specification of the formant peaks, or there may be sufficient cues remaining in vowels with high f_0 to fill in the details lost by sparse sampling. Alternative models have been proposed that do not rely on formants, but instead use spectral shape parameters derived from principal components analysis or related methods (e.g., Bakkum *et al.*, 1995; Zahorian and Jaghargi, 1993). A question for future research is how these models would account for the effects of formant flattening and their independence of changes in voicing source. The auditory excitation patterns that generally serve as the input to spectral shape models assign greater weight to the low-frequency region of the spectrum, and are therefore likely to show increased, rather than reduced, dependence on f_0 .

A second question concerns the basis for the beneficial

effects of time-varying spectral change: if not by improving spectral resolution, how then does formant movement help? The answer may be that formant movement provides a basis for maintaining phonetic distinctiveness in crowded vowel systems such as English (Ladefoged and Maddieson, 1996). Several of our findings are consistent with this hypothesis. First, the effects of formant-frequency flattening were not uniform across the vowel set. Consistent with previous investigations demonstrating the perceptual salience of time-varying formant frequencies for American English vowels (e.g., Strange *et al.*, 1983; Nearey and Assmann, 1986; Hillenbrand, 1995), the largest decrements in identification were found for the vowels that showed the largest excursions in their formant trajectories, /e/ and /o/. Second, there was a tendency for neighboring vowels in the $F1$ - $F2$ plane to display opposing patterns of formant movement (Nearey and Assmann, 1986). This can be seen by inspection of Fig. 3: vowels whose onsets are close together (such as /e/ and /ε/) tend to display vector movement in opposing directions. This suggests that formant movement provides greater benefits in crowded regions of the vowel space. If formant movement serves a contrast-enhancing function in vowel perception, then manipulations such as flattening the formants might be predicted to have reduced effects in languages with smaller vowel inventories.

ACKNOWLEDGMENTS

This research was supported by Grant No. 11423-590 from the Texas Advanced Research Program, awarded by the Texas Higher Education Coordinating Board. Portions of this research were reported at the 130th Meeting of the Acoustical Society of America in St. Louis, MO. The authors would like to thank Kathleen Jenouri, Matt Sommer, Elaine Teoh, Ginger Stickney, Debbie Moncrieff, Phillip Hamilton, and Charles Rees for their assistance in data collection and analysis, Terry Nearey for help in software development and thoughtful discussions, and Chris Darwin, Keith Kluender, and an anonymous reviewer for helpful comments on the manuscript.

¹This prediction rests on the assumption that the detrimental effects of higher f_0 are not accompanied by other differences between the vowels of adults and children that lead to *improved* resolution of children's vowels compared to adults, such as narrower formant bandwidths or increased frequency separation between adjacent formants. This caveat is discussed in Sec. II I 3.

²The spectra of voiced vowels contain peaks at the frequencies of the harmonics when the temporal window is longer than the period of the fundamental. When f_0 is high, these peaks may be too sparse to allow for accurate formant estimation. However, the harmonic structure can be eliminated by performing a spectral analysis with a temporal window equal to or shorter than the duration of a single pitch period. An accurate representation of the vocal tract transfer function can be obtained from a pitch-synchronous analysis, but this approach requires precise estimates of f_0 (de Cheveigné and Kawahara, 1999).

³One of the reviewers raised the concern that the initial sample used to derive the formant parameters for the flattened-formant conditions was taken somewhat later in the vowel than in earlier studies, and therefore may have underestimated the formant movement in the vowels. Hillenbrand and Nearey (1999) used an analysis window centered at 20% and 80% of the duration of the vocalic nucleus. We repeated the acoustic analysis using their 20%–80% time window and obtained the same statistical results as reported in Sec. II D above, although several vowels did show increased

formant movement with more extreme time windows (particularly /*ʌ*/, /*æ*/, /*e*/, and /*o*/ in *F1*, and /*ɔ*/, /*a*/, /*u*/, and /*ʌ*/ in *F2*). The perceptual consequences of selecting a later sample are probably minor, however, as a subset of these vowels with reduced formant movement when measured with the 33%–66% time window (e.g., /*ʌ*/, /*æ*/, /*a*/, /*u*/) showed little decline in identification accuracy from the full to FlatF conditions of experiment 1 (see Fig. 6).

⁴This interpretation must be treated with caution because Lee *et al.* (1999) reported greater time-varying spectral change for children's vowels compared to those of adults. Our acoustic analyses provided little evidence of increased time variation in the formant trajectories of children's vowels. Differences in vowel set, elicitation procedure, and analysis method may be responsible for the discrepancy.

- Andruski, J. E., and Nearey, T. M. (1992). "On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables," *J. Acoust. Soc. Am.* **91**, 390–410.
- Assmann, P. F., and Nearey, T. M. (1987). "Perception of front vowels: The role of harmonics in the first formant region," *J. Acoust. Soc. Am.* **81**, 520–534.
- Assmann, P. F., Ballard, W., Bornstein, L., and Paschall, D. D. (1994). "TrackDraw: A graphical interface for controlling the parameters of a speech synthesizer," *Behav. Res. Methods Instrum. Comput.* **26**, 431–436.
- Bakkum, M. J., Plomp, R., and Pols, L. W. (1995). "Objective analysis versus subjective assessment of vowels pronounced by deaf and normal-hearing children," *J. Acoust. Soc. Am.* **98**, 745–762.
- Darwin, C. J., and Gardner, R. B. (1985). "Which harmonics contribute to the estimation of first-formant frequency?" *Speech Commun.* **4**, 231–235.
- de Cheveigné, A., and Kawahara, H. (1999). "Missing data model of vowel perception," *J. Acoust. Soc. Am.* **105**, 3497–3508.
- Delattre, P. C., Liberman, A. M., and Cooper, F. S. (1955). "Acoustic loci and transitional cues for consonants," *J. Acoust. Soc. Am.* **27**, 769–773.
- Di Benedetto, M. G. (1989). "Frequency and time variations of the first formant: Properties relevant to the perception of vowel height," *J. Acoust. Soc. Am.* **86**, 67–77.
- Diehl, R. L., Lindblom, B., Hoemeke, K. A., and Fahey, R. P. (1996). "On explaining certain male-female differences in the phonetic realization of vowel categories," *J. Phonetics* **24**, 187–208.
- Dissard, P., and Darwin, C. J. (2000). "Extracting spectral envelopes: formant frequency matching between sounds on different and modulated fundamental frequencies," *J. Acoust. Soc. Am.* **107**, 960–969.
- Eguchi, S., and Hirsh, I. J. (1969). "Development of speech sounds in children," *Acta Oto-Laryngol. Suppl.* **257**, 1–51.
- Fox, R. A. (1989). "Dynamic information in the identification and discrimination of vowels," *Phonetica* **46**, 97–116.
- Fujimura, O., and Lindqvist, J. (1971). "Sweep-tone measurements of vocal-tract characteristics," *J. Acoust. Soc. Am.* **49**, 541–558.
- Hawks, J. W., and Miller, J. D. (1995). "A formant bandwidth estimation procedure for vowel synthesis," *J. Acoust. Soc. Am.* **97**, 1343–1344.
- Hillenbrand, J. (1995). "Identification of vowels resynthesized from /hVd/ utterances: Effects of formant contour," *J. Acoust. Soc. Am.* **97**, 3245(A).
- Hillenbrand, J., and Gayvert, R. T. (1993). "Identification of steady-state vowels synthesized from the Peterson and Barney measurements," *J. Acoust. Soc. Am.* **94**, 668–674.
- Hillenbrand, J., and Nearey, T. M. (1999). "Identification of resynthesized /hVd/ utterances: Effects of formant contour," *J. Acoust. Soc. Am.* **105**, 3509–3523.
- Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Am.* **97**, 3099–3111.
- Katz, W. F., and Assmann, P. F. (2000). "Identification of children's and adult's vowels: Intrinsic fundamental frequency, presence of voicing, and voicing dynamics," submitted to *J. Phonetics*.
- Kent, R. D. (1976). "Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic study," *J. Speech Hear. Res.* **19**, 421–445.
- Kewley-Port, D., and Watson, C. S. (1994). "Formant-frequency discrimination for isolated English vowels," *J. Acoust. Soc. Am.* **95**, 485–496.
- Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.* **67**, 971–995.
- Ladefoged, P., and Maddieson, I. (1996). *The Sounds of the World's Languages* (Blackwell, London).
- Lee, S., Potamianos, A., and Narayanan, S. (1999). "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *J. Acoust. Soc. Am.* **105**, 1455–1468.
- Meddis, R., and Hewitt, M. (1991). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," *J. Acoust. Soc. Am.* **89**, 2866–2882.
- Moore, B. C. J., and Glasberg, B. R. (1987). "Formulas describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns," *Hear. Res.* **28**, 209–225.
- Nábělek, A. K., and Ovchinnikov, A. (1997). "Perception of nonlinear and linear formant trajectories," *J. Acoust. Soc. Am.* **101**, 488–497.
- Nearey, T. M. (1989). "Static, dynamic, and relational properties in vowel perception," *J. Acoust. Soc. Am.* **85**, 2088–2113.
- Nearey, T. M., and Assmann, P. F. (1986). "Modelling the role of inherent spectral change in vowel identification," *J. Acoust. Soc. Am.* **80**, 1297–1308.
- Peterson and Barney. (1952). "Control methods used in a study of vowels," *J. Acoust. Soc. Am.* **24**, 175–184.
- Pols, L. C. W., and Van Son, R. J. H. (1993). "Acoustics and perception of dynamic vowel segments," *Speech Commun.* **13**, 135–147.
- Ryalls, J., and Lieberman, P. (1982). "Fundamental frequency and vowel perception," *J. Acoust. Soc. Am.* **72**, 1631–1634.
- Rosner, B. S., and Pickering, J. B. (1994). *Vowel Perception and Production* (Oxford University Press, New York).
- Smith, B. L., Kenney, M. K., and Hussain, S. (1995). "A longitudinal investigation of duration and temporal variability in children's speech production," *J. Acoust. Soc. Am.* **99**, 2344–2349.
- Stevens, K. H. (1999). *Acoustic Phonetics* (M.I.T. Press, Cambridge, MA).
- Strange, W. (1989). "Evolving theories of vowel perception," *J. Acoust. Soc. Am.* **85**, 2081–2087.
- Strange, W., Jenkins, J. J., and Johnson, T. L. (1983). "Dynamic specification of coarticulated vowels," *J. Acoust. Soc. Am.* **74**, 695–705.
- Sussman, H. M., McCaffrey, H. A. L., and Matthews, S. A. (1991). "An investigation of locus equations as a source of relational invariance for stop place categorization," *J. Acoust. Soc. Am.* **90**, 1309–1325.
- Tartter, V. C. (1991). "Identifiability of vowels and speakers from whispered syllables," *Percept. Psychophys.* **49**, 365–372.
- Zahorian, S. A., and Jaghargi, A. J. (1991). "Speaker normalization of static and dynamic vowel spectral features," *J. Acoust. Soc. Am.* **90**, 67–75.
- Zahorian, S. A., and Jaghargi, A. J. (1993). "Spectral-shape features versus formants as acoustic correlates for vowels," *J. Acoust. Soc. Am.* **94**, 1966–1982.