Acoustics of children's speech: Developmental changes of temporal and spectral parameters[a)]

Sungbok Lee,[b)] Alexandros Potamianos, and Shrikanth Narayanan
AT&T Labs–Research, 180 Park Avenue, Florham Park, New Jersey 07932-0971

(Received 11 December 1997; revised 30 September 1998; accepted 3 November 1998)

Changes in magnitude and variability of duration, fundamental frequency, formant frequencies, and spectral envelope of children's speech are investigated as a function of age and gender using data obtained from 436 children, ages 5 to 17 years, and 56 adults. The results confirm that the reduction in magnitude and within-subject variability of both temporal and spectral acoustic parameters with age is a major trend associated with speech development in normal children. Between ages 9 and 12, both magnitude and variability of segmental durations decrease significantly and rapidly, converging to adult levels around age 12. Within-subject fundamental frequency and formant-frequency variability, however, may reach adult range about 2 or 3 years later. Differentiation of male and female fundamental frequency and formant frequency patterns begins at around age 11, becoming fully established around age 15. During that time period, changes in vowel formant frequencies of male speakers is approximately linear with age, while such a linear trend is less obvious for female speakers. These results support the hypothesis of uniform axial growth of the vocal tract for male speakers. The study also shows evidence for an apparent *overshoot* in acoustic parameter values, somewhere between ages 13 and 15, before converging to the canonical levels for adults. For instance, teenagers around age 14 differ from adults in that, on average, they show shorter segmental durations and exhibit less within-subject variability in durations, fundamental frequency, and spectral envelope measures. [S0001-4966(99)03202-6]

PACS numbers: 43.10.Ln, 43.70.Ep [AL]

INTRODUCTION

In speech-development research, it is important to know how acoustic parameters of speech such as fundamental frequency, formant frequencies, and segmental durations vary as a function of age and gender, and at what age the magnitude and variability of acoustic parameters begin to exhibit adult-like patterns. When properly interpreted, such chronological knowledge of speech acoustics could provide insights into the underlying development of speech organs and speech-motor control in children, and help in creating an accurate developmental model of the vocal tract (cf. Goldstein, 1980). Previous studies have shown that children's speech, compared to adults' speech, exhibits higher pitch and formant frequencies, longer segmental durations, and greater temporal and spectral variability (Eguchi and Hirsh, 1969; Kent, 1976; Kent and Forner, 1980; Smith, 1978, 1992; Smith *et al.*, 1995; Hillenbrand *et al.*, 1995). However, due to inadequate data in terms of either the total number of subjects or the age range of the subjects, these studies provide only limited information on the acoustic patterns of speech from early childhood through adulthood. A more chronologically detailed acoustic database obtained from a larger number of subjects with a wider age range is needed in order to better understand developmental acoustic patterns in children's speech and their relation to the underlying anatomical and neuromuscular development. This study represents an effort in that direction.

The study is also motivated by speech applications such as automatic speech/speaker recognition and speech synthesis. In recent years, the problem of automatic recognition of children's speech has gained attention (Palethorpe *et al.*, 1996; Potamianos *et al.*, 1997). Potamianos *et al.* (1997) have shown that the performance of a hidden Markov model (HMM) speech-recognition system trained on adults' speech degrades substantially when tested on the speech of children age 12 and younger. In addition to the acoustic differences between children's and adults' speech, the acoustic variability inherent in children's speech contributes to the degradation in recognition performance as demonstrated by the results of Palethorpe *et al.* (1996): Performance in classification of vowels produced by children of 5 years of age is much worse than for that of adults (60%–65% vowel-classification accuracy for children versus over 90% for adults). Chronologically detailed acoustic data obtained from a large number of speakers can be helpful in devising strategies for dealing with the acoustic mismatch between different age groups.

This paper reports on a set of temporal and acoustic parameters measured from a speech database recently collected from 436 subjects ages 5 through 18 and from 56 adults (Miller *et al.*, 1996). For vowels, magnitude and variability of durations, fundamental frequency ($F0$), and the first three formant frequencies ($F1-F3$) as well as spectral-envelope variability are measured and analyzed as a function of age and gender. Duration magnitude and variability are

[a)]''Selected research articles'' are ones chosen occasionally by the Editor-in-Chief that are judged (a) to have a subject of wide acoustical interest, and (b) to be written for understanding by broad acoustical readership.

[b)]Present address: Lucent Technologies–Bell Labs, 700 Mountain Avenue, Murray Hill, NJ 07974.

TABLE I. Distribution of subjects by age (in years) and gender.

Age	5	6	7	8	9	10	11	12	13	14	15	16	17	18	5–18	25–50
Male	19	11	11	25	23	25	24	22	16	11	11	11	10	10	229	29
Female	13	16	24	11	25	14	19	21	13	10	11	11	9	10	207	27
Total	32	27	35	36	48	39	43	43	29	21	22	22	19	20	436	56

also measured for the fricative /s/ as well as for the recitation of sentence-length utterances. Results and findings are presented with a focus on age- and gender-dependent acoustic changes occurring during the course of speech development. The paper is organized as follows: In Sec. I, the database used in the current study is described. In Sec. II, procedures used for acoustic measurements and statistical analyses are described. Results are presented in Sec. III, followed by a discussion in Sec. IV.

I. SPEECH DATABASE

The database analyzed in this study was collected from 436 children, ages 5 through 18 with a resolution of 1 year of age, and from 56 adult speakers (ages 25–50). The data collection was a joint effort of Southwestern Bell Technology Resources and the Central Institute for the Deaf (CID), and was carried out over a period of approximately 6 months. The recording site was located in the St. Louis Science Center (Missouri), a popular attraction among children, which enabled easy recruitment of subjects. The distribution of subjects by age and gender is shown in Table I. Among the 492 subjects, 316 were born and raised in the two Midwestern states of Missouri and Illinois.

The speech material consisted of ten monophthongal and five diphthongal vowels in American English and five phonetically rich meaningful sentences (the diphthongs are not analyzed in this paper). Target words for the ten monophthongs analyzed in this study were *bead* (/IY/), *bit* (/IH/), *bet* (/EH/), *bat* (/AE/), *pot* (/AA/), *ball* (/AO/), *but* (/AH/), *put* (/UH/), *boot* (/UW/), and *bird* (/ER/). The target words were produced in the carrier sentence ''I say uh --- again'' except for children of ages 5 and 6, who produced target words in isolation. A schwa-like sound [uh] was used as an attempt to make subjects maintain a neutral vocal tract before producing target words. The five sentences were: (1) ''He has a blue pen.'' (2) ''I am tall.'' (3) ''She needs strawberry jam on her toast.'' (4) ''Chuck seems thirsty after the race.'' (5) ''Did you like the zoo this spring?''

Recordings were made in a sound-treated booth located inside a glass-panel enclosure, using a high-fidelity microphone (Bruel & Kjaer model #4179) connected to a real-time waveform digitizer with 20-kHz sampling rate and 16-bit resolution. The target utterances were presented on a computer monitor *twice* in random order. No specific instructions were given to the subjects regarding the manner of production. Prior to the recording session, any target utterances that the speakers (mostly 5- and 6-year-olds) had difficulty reading were identified and elicited through imitation of a sample prerecorded by a female speech pathologist.

After the data collection, each waveform file was manually examined by listening to the recorded speech. Waveform files that were truncated or of very low recording quality were marked as ''chopped'' or ''bad'' and are excluded from this study. From the initial 24 630 waveform files, 24 152 files were judged to be good and were included in the database.

II. SPEECH ANALYSIS

A. Preprocessing of the database

In order to process the large number of speech samples, an automatic procedure was utilized for the necessary phoneme-level segmentation of each utterance. The AT&T hidden Markov model recognition engine (Ljolje and Riley, 1991) was used for the purpose. Specifically, a set of hidden Markov models (HMMs) of phonemes trained from adult speakers was used to obtain initial phonemic segmentation of the present children's database. Next, in order to minimize the initial phonemic-alignment error due to acoustic differences between adults' and children's speech, the initially segmented children's speech was used to retrain the HMMs. Finally, the database was resegmented using the retrained HMMs.

For each utterance, the automatic segmentation procedure produced a label file in which the beginning and the end of each phoneme and pause period were time-marked, according to the entry time to the first state and the exit time from the last state, respectively, of the corresponding HMM phoneme unit. Time marks have a 10-ms resolution, or uncertainty, which is half the length of the analysis window.

In order to examine the accuracy of the automatic segmentation procedure, durations of 160 vowel utterances from 16 randomly selected subjects of ages 5, 8, 11, and adults were manually measured. Mean segmentation difference between the automatically computed and the manually measured values was -17.5 ms with a standard deviation of 37.0 ms. As indicated by the negative mean difference, vowel durations were somewhat underestimated by the automatic segmentation procedure. However, no appreciable age-dependent trend was found in the mean segmentation difference across the four age groups investigated. Hence, age-dependent duration *trends* in the postsegmentation data are preserved even though duration *values* may be somewhat smaller. It should however be noted that second-order statistics of duration measurements will be noisy (as indicated by the high variance in the manual versus automatic segmentation differences) and should be interpreted with caution. Erroneous segmentation will have minimal effect on the statistics of $F0$ and formant frequency values since comparative

analyses used global median values for each token, as discussed in the next section.

B. Duration

Durations of the ten monophthongal vowels, the fricative portion of /s/ in the carrier sentence (''I [s]ay−''), and the five sentences were measured from the corresponding label files produced by the automatic segmentation procedure. Since each of the sentences was of different length, the sentence durations were normalized with respect to the corresponding mean sentence duration of adult male speakers.

Since the automatic segmentation procedure sometimes erroneously yielded excessively short or long vowel and /s/ durations, a crude effort was made to minimize the inclusion of such outliers using duration histograms. First, vowels with duration less than 80 ms were discarded along with their pair (i.e., repetitions of the same vowel by the same subject). Among the initial 9424 tokens on which durations were successfully measured, 4404 vowel pairs (8808 tokens) were included in the data set. It was noted that correlation of durations between the first and second productions was somewhat weak ($r=0.69$). Next, when difference in duration between two repetitions of the same vowel by the same speaker was greater than 160 ms, that vowel pair was also discarded. This yielded 3793 vowel pairs (7586 tokens) with substantially enhanced correlation ($r=0.83$). These 3793 vowels pairs are analyzed in the current study.[1] In the case of /s/, tokens with duration smaller than 90 ms or larger than 250 ms were discarded. Further, only /s/ tokens collected from subjects with at least 30 repetitions were included in the analysis. Out of the initial 16 897 /s/ tokens from 431 subjects, 15 592 tokens from 396 subjects were included in the final data set.

Duration data were organized by vowel, age, and gender group and analyzed using the SPSS statistical software package. Between- and within-subject variability and group means were estimated and group mean comparisons were performed.

C. Fundamental and formant frequencies

The fundamental frequency ($F0$) and the first three formant frequencies ($F1-F3$) of the ten monophthongs were estimated using the automatic $F0$ and formant-tracking program in the ESPS signal-processing package by Entropic Research Laboratory. The software utilizes a dynamic programming technique to select the best pitch and formant tracks from raw pitch and formant estimates (Secrest and Doddington, 1983). Each speech waveform was downsampled to 12 kHz and processed using a 12-ms Hamming window with 5-ms window update, a first-difference pre-emphasis factor of 0.94, and a 12th-order linear-prediction analysis. The resulting raw $F0$ and formant tracks were smoothed using a 3-point median filter. Global median values of the entire tracks were computed and used as the representative $F0$ and formant frequencies for the tracks.

The performance of the automatic program was evaluated using hand-measured pitch and formant values of 96 randomly selected vocalic segments from 16 subjects ages 5, 8, 11, and adults. The manual estimation was done as follows: (1) three 20-ms segments were selected around the steady-state portion of each vocalic segment, (2) the pitch and formants of each segment were measured by visual inspection of the corresponding discrete Fourier transform (DFT) spectrum, spectrogram, and the locations of formant peaks in the linear predictive (LP) spectral envelope, and (3) the hand-measured values of pitch and formants were averaged over the three subsegments. The manually estimated $F0$ and formant frequencies were compared to the automatically computed ones. Mean differences (standard deviation) between the automatically and manually estimated values in Hz were 7.6 (23.8) for $F0$, 43.6 (87.2) for $F1$, 92.4 (183.8) for $F2$, and 193.1 (400.7) for $F3$.

The automatic formant-tracking program yielded reasonable estimates of the first formant frequency in most cases. $F2$ and $F3$, however, were often inaccurate for vowels produced by young children due to poor spectral resolution at high frequencies (partially caused by wider harmonic spacing and breathy voicing), spurious spectral peaks, and formant-track merging. In such cases, manual estimation of formants from the speech spectrogram was also difficult. Therefore, in order to minimize statistical biases due to erroneously estimated formant frequencies, the raw formant data were refined using the following procedure: the initial formant data of all subjects were grouped according to vowel, age, and gender (10 vowels$\times$15 age groups$\times$2 genders). Next, two two-sigma ellipses were computed from ($F1$,$F2$) and ($F2$,$F3$) data sets, respectively, and data points that fell outside the region of either ellipse were removed. After the removal of the outliers, the mean and standard deviation were computed and each data file was visually examined: Whenever one of the $F1-F3$ values was subjectively judged to be too low or too high, the corresponding formant set was discarded. From the initial set of 9424 vowel tokens, 7631 tokens of the first three formant frequencies were included in the final data set analyzed in this study.[2] Despite our efforts to remove erroneous formant values, it is possible that the refined formant-data set still includes some underestimated $F2$ and $F3$ values, especially for children ages 7 and lower.

Pitch tracking by the automatic program was fairly reliable across age and gender except for occasional pitch-halving. The large s.d. (23.8 Hz) of the mean difference between the automatically and manually measured $F0$ data is mostly due to such occurrences. Inclusion of such erroneous $F0$ data was minimized using the two-sigma ellipse method described above applied to individual $F0$ data.

The $F0$ and formant data sets were organized by vowel, age, and gender and analyzed using the SPSS software package. Group means, between- and within-subject variations were estimated. For the computation of within-subject variation of formant frequencies, formant data of only matched vowel pairs (3265 pairs) were considered. In addition, between- and within-subject coefficients of variation (COV) were computed by taking the ratio of s.d. to the corresponding mean. The COV has been used by Eguchi and Hirsh (1969) to minimize a possible positive correlation between magnitude and variability.

D. Spectral-envelope variability

Spectral distance measures using a set of cepstrum coefficients derived from a log-spectral envelope representation are widely used in automatic speech recognition (Rabiner and Juang, 1993). In this paper, two spectral-envelope variability measures are computed from the cepstrum coefficients: (1) spectral distance between two repetitions of the same target vowel, and (2) spectral distance between the first- and second-half portions of vocalic segments including transitional regions from (to) the preceding (following) consonants. The former is compared with the within-subject formant variability in order to test whether the age-dependent reduction of the formant variability is a general phenomenon associated with speech development (Eguchi and Hirsh, 1969). The latter is interpreted as a measure of spectral movement, or transition, inside a vowel. An implicit assumption is that a greater difference in the underlying articulatory configuration between two vocalic segments induces a greater distance between the corresponding spectral envelopes.

The spectral envelope of a given vocalic segment was computed using a mel-frequency filterbank of 29 frequency bands spanning 100 to 6000 Hz (Davis and Mermelstein, 1980). The 12 cepstrum coefficients were computed from the log-spectral envelope using the inverse cosine transform. The distance between two speech segments was computed by first computing the average short-time cepstrum vector for each of the segments using a 20-ms window, and then taking the Euclidean distance between the two average cepstrum vectors. The zeroth-order cepstral coefficient (energy term) was not considered in the spectral distance computation since it does not affect the shape of the spectral envelope.

III. RESULTS

A. Phoneme and sentence durations

1. Vowel durations

Since it was found that the effect of gender on vowel duration is not significant, durations averaged across all vowels and subjects in each age group are shown in Fig. 1(a). Error bars denote between-subject variations. Effect of age is significant [$F(14,3778)=36.2$, $p<0.005$].

Multiple *a priori* comparisons (Bonferroni test with significance level 0.05) show that the groups of age 5 (279 ms) and age 6 (264 ms) exhibit significantly longer averaged vowel durations than older age groups. This may be partially due to the difference in the mode of speech elicitation, since the two age groups produced target words in isolation. The multiple comparisons also show that the reduction of vowel duration from age 10 (199 ms) to age 12 (178 ms) and from age 11 (191 ms) to age 15 (168 ms) are significant, while no significant difference in vowel duration exists among age groups older than 12. On average, vowel durations reach minima around age 15. Furthermore, difference in mean duration between age 15 and adults is significant ($t=2.42$, $df=644$, $p<0.02$). Therefore, it is possible that vowel durations increase again in the process of converging toward adult range. As will be shown later, similar trends are also observable for both /s/ and sentence durations as well as for some spectral parameters.

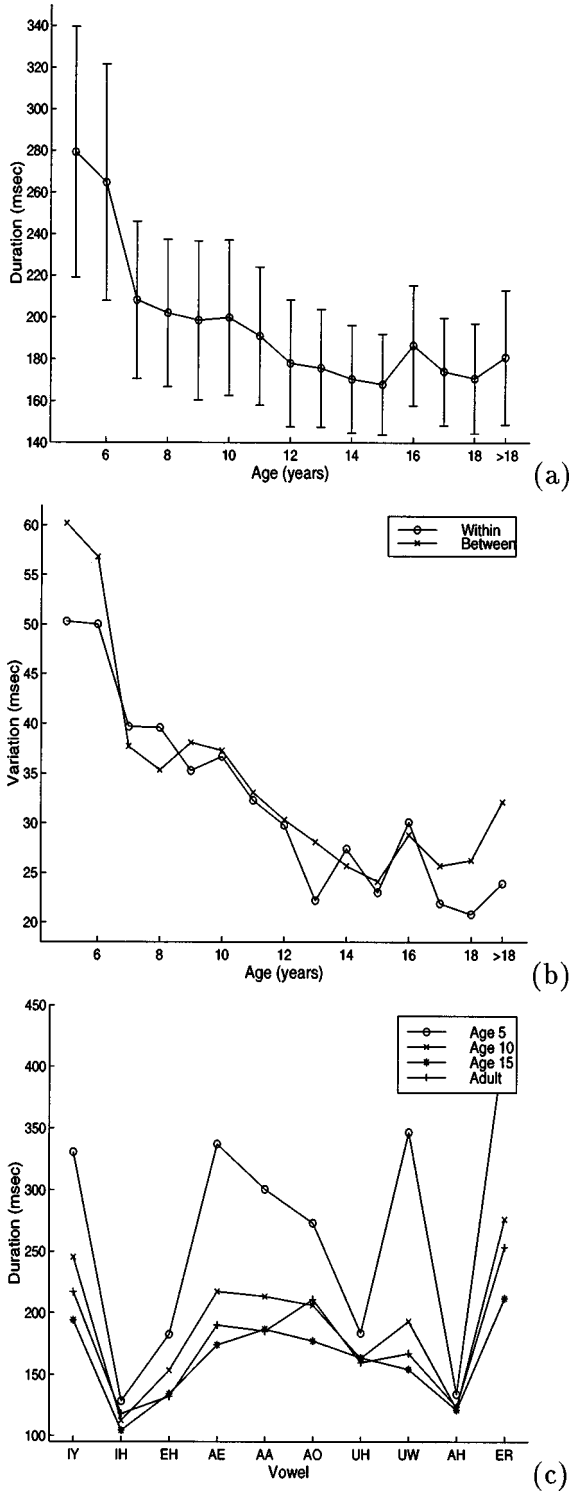

FIG. 1. (a) Averaged-vowel duration across all vowels and subjects in each age group. (b) Within- and between-subject variations. The between-subject variation is reduced by a factor of 2.0. (c) Mean duration of individual vowel averaged across all subjects in each age group is shown for several age groups.

In Fig. 1(b), within- and between-subject variations are shown as a function of age. The between-subject variation shown is reduced by a factor of 2.0 in order to facilitate the

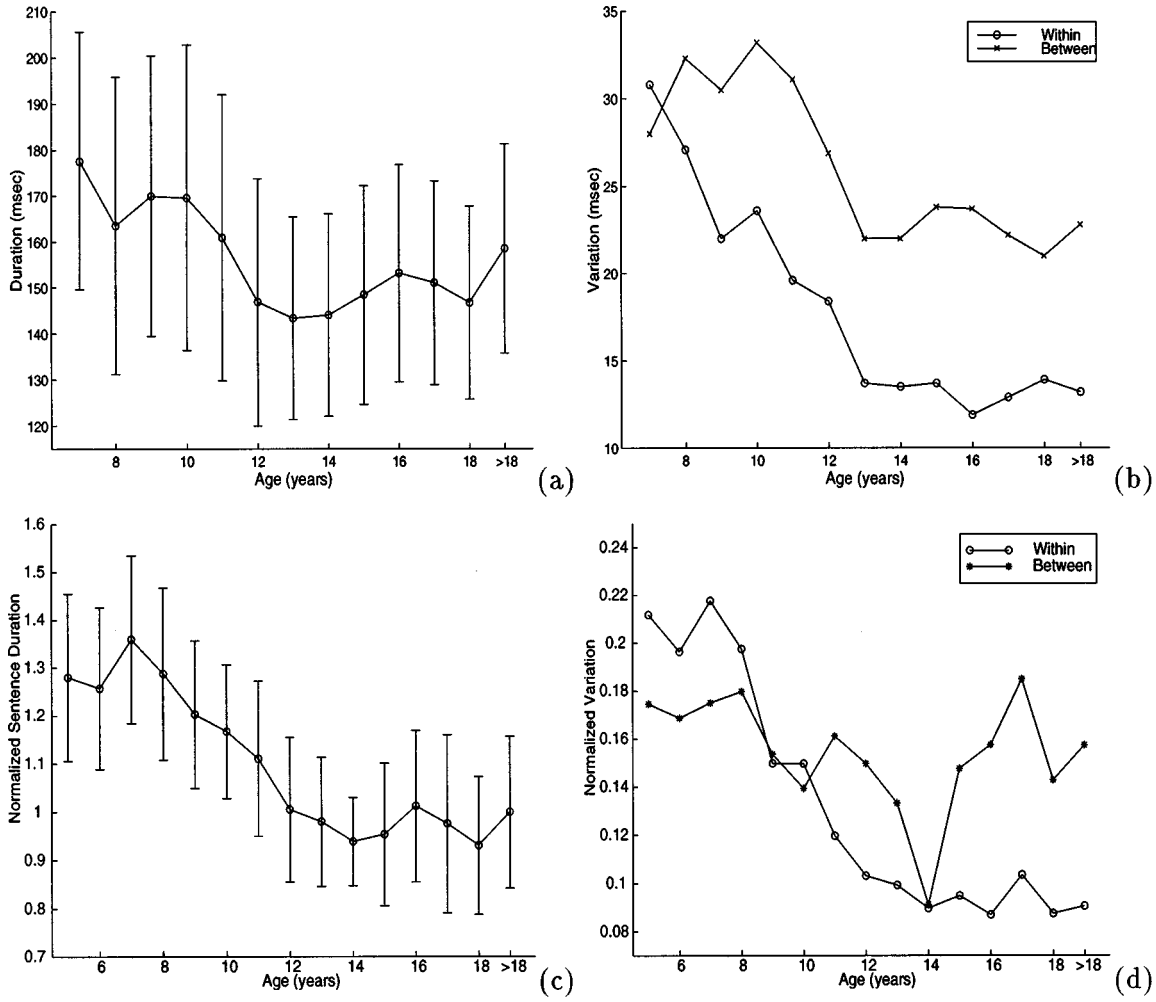

FIG. 2. (a) Duration of /s/. (b) Within- and between-subject variability of /s/ duration (unfilled circle: within-subject, filled circle: between-subject). (c) Normalized duration of sentence. (d) Variability of sentence duration (unfilled circle: within-subject, filled circle: between-subject).

comparison of within- and between-subject variability. It is clear that both within- and between-subject variations decrease with age. For the within-subject variation, the effect of age is significant [$F(14,3778)=18.5$, $p<0.001$]. The multiple comparisons indicate that the reduction of within-subject variation from age 11 to age 13 is significant ($p<0.05$), while no significant difference exists among the age groups of 13 and older. Therefore, both duration and within-subject variability may reach adult level almost simultaneously around age 12. Also note the remarkably similar behavior of between-subject variability and duration.

Individual vowel durations averaged across subjects in each age group are shown in Fig. 1(c) for several age groups. The similar vowel duration patterns between 5-year-olds and adults suggest that children as young as 5 years old have the ability to control intrinsic vowel duration. However, 5-year-olds show a tendency to exaggerate the duration of ''long'' vowels, e.g., /IY/, /AE/, /AA/, /ER/, when compared to the other age groups. The fact that children of ages 5 and 6 produced vowels in isolation may be at least partially responsible for this trend.

2. /s/ and sentence production durations

Results of duration and variability measurements are shown in Fig. 2(a) and (b) for /s/ and in Fig. 2(c) and (d) for sentence productions as a function of age. Since the effect of gender is not significant, values averaged across gender are shown in the figure. However, the effect of age is significant [$F(12,383)=4.178$, $p<0.001$ for /s/; $F(14,469)=23.1$, $p<0.001$ for sentences]. Note that in Fig. 2(a) and (b) no data points are shown for ages 5 and 6 since target words were produced in isolation (no /s/ tokens).

The multiple comparisons indicate that the duration of /s/ significantly decreases from age 10 (170 ms) to age 12 (147 ms). On average, the duration of /s/ reaches minima around age 13 and increases again toward adult levels. Adults' mean duration (159 ms) is significantly different from that of age 13 (143 ms) ($t=3.02$, $df=83$, $p<0.005$).

The within- and between-subject variability of /s/ [Fig. 2(b)] decrease gradually up to age 13 and then remain more or less constant. The effect of age on the within-subject variation is significant [$F(12,383)=19.3$, $p<0.001$]. Multiple comparisons indicate that significant reduction of within-subject variation occurs from age 10 (23.6 ms) to age 12 (18.4 ms) and from age 11 (19.6 ms) to age 13 (13.7 ms). After age 13, no significant change of the within-subject variation is observed. Therefore, the within-subject variability reaches adult level around age 13. Note that the significant increase of /s/ duration from age 13 (143 ms) to adult

(159 ms) ($t=3.01$, $df=82$, $p<0.03$) is not accompanied by a similar increase in within-subject variability. In addition, the sudden and substantial reduction in between-subject variability from age 11 to age 13 is worth noticing.

It is likely that sentence duration shown in Fig. 2(c) is mainly determined by the speaking rate, reading ability, and pause duration. The effect of age on sentence duration is significant [$F(14,481)=23.1$, $p<0.001$]. It is observed that sentence duration decreases almost linearly from age 7 to age 14, where it attains its minimum value. About a 45% reduction in duration occurs in that time period. The multiple comparisons indicate that the reduction of duration from age 10 to age 12 and from age 11 to age 14 are significant. Mean durations at age 14 and for adults are also significantly different ($t=2.17$, $df=75$, $p<0.05$). Therefore, just as in the case of vowels and /s/, sentence durations also reach a minimum before converging toward adult levels.

Note that the relatively short average-sentence duration for 5- and 6-year-old children [Fig. 2(c)] was not due to measurement errors or the different elicitation method used for some young children with reading problems (sentence durations were very similar for both ''repeat after me'' and ''read'' elicitation methods). From listening to the sentence productions of young children, it was found that disfluencies (phoneme deletions, mumbling of groups of sounds, mispronunciations) were quite common, and are suspected to be the main cause for shorter sentence durations, particularly in 5- and 6-year-olds.

The effect of age on the within-subject variability is significant [$F(14,477)=14.5$, $p<0.001$]. Multiple comparisons indicate that the reduction of within-subject variations from age 8 to age 12 is significant, while there is no significant change after age 12. An unexpected and interesting observation is the large between-subject variation for subjects older than age 14.

B. Fundamental frequency

The mean $F0$ of male and female speakers averaged across all vowels and subjects in each age group is shown in Fig. 3(a) as a function of age. Vertical bars denote between-subject variations. The mean $F0$ values for male and female speakers, averaged across all subjects for each vowel and age group, are provided in Tables II and III. A simple factorial analysis of variance (ANOVA) indicates that $F0$ differences between male and female speakers become significant beginning from age 12. For male speakers, multiple comparisons indicate that the drops in $F0$ from age 11 to age 13 and from age 13 to age 15 are significant ($p<0.05$). About a 78% drop in $F0$ occurs between age 12 ($F0=226$ Hz) and age 15 ($F0=127$ Hz) in male speakers, and there is no significant pitch change after age 15. This suggests that, on average, pubertal pitch change in male speakers starts between age 12 and 13, and ends around age 15. The relatively large between-subject variation at ages 13 and 14 also suggests that the onset time of puberty is different among speakers in these age groups (cf. Hollien *et al.*, 1994). For female speakers, multiple comparisons indicate that the pitch drop from age 7 ($F0=275$ Hz) to age 12 ($F0=231$ Hz) is significant,

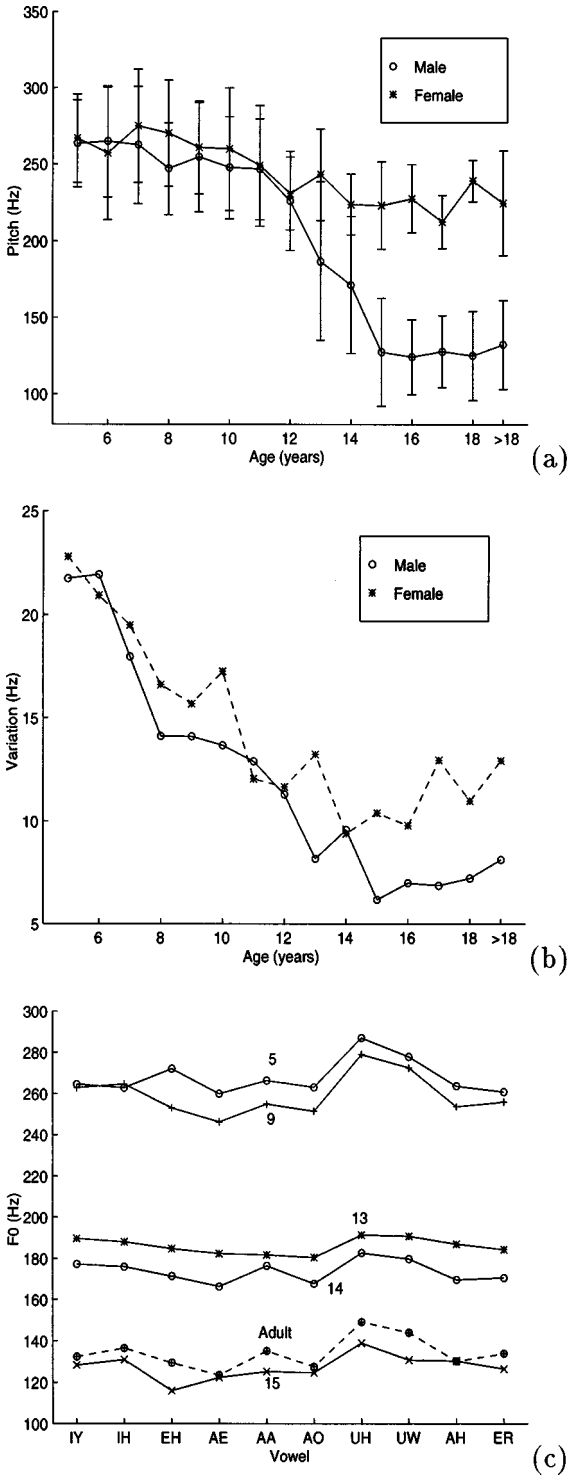

FIG. 3. (a) Averaged fundamental frequency for male and female speakers. Vertical bars denote between-subject variations (i.e., group standard deviations). (b) Within-subject pitch variation as a function of age and gender. (c) Mean pitch of individual vowels for male speaker as a function of age.

and there is no significant pitch change after that age. The $F0$ change for female subjects is more gradual compared to male speakers.

The within-subject variation of pitch is shown in Fig. 3(b) as a function of age and gender. The effect of age is significant for both genders [$F(14,242)=11.9$, $p<0.001$ for male speakers; $F(14,217)=6.0$, $p<0.001$ for female speakers]. For male speakers, the multiple comparisons indicate

TABLE II. Mean and standard deviation (in parentheses) of fundamental frequency and formant-frequency values for male speakers (# is the number of tokens).

Age		aa	ae	ah	ao	eh	er	ih	iy	uh	uw
	#	26	20	25	28	22	26	23	18	26	22
	*F*0	266 (33)	260 (37)	263 (31)	263 (35)	272 (34)	260 (35)	262 (35)	264 (32)	287 (42)	277 (35)
5	*F*1	1166 (98)	1010 (117)	824 (129)	837 (78)	835 (101)	620 (70)	636 (74)	467 (67)	656 (59)	477 (45)
	*F*2	1750 (165)	2534 (163)	1669 (195)	1226 (99)	2475 (216)	1705 (154)	2608 (171)	3071 (145)	1434 (98)	1508 (285)
	*F*3	3412 (370)	3452 (252)	3596 (300)	3533 (268)	3469 (321)	2502 (353)	3465 (307)	3653 (215)	3588 (353)	3136 (294)
	#	21	18	18	19	17	15	14	14	18	18
	*F*0	273 (49)	256 (40)	263 (35)	264 (39)	259 (34)	257 (39)	261 (34)	267 (30)	290 (40)	279 (42)
6	*F*1	1048 (88)	962 (84)	844 (75)	813 (84)	888 (45)	661 (41)	652 (48)	411 (83)	663 (33)	457 (60)
	*F*2	1554 (117)	2544 (136)	1534 (107)	1147 (117)	2506 (107)	1565 (180)	2728 (141)	3124 (161)	1363 (120)	1659 (270)
	*F*3	3292 (322)	3316 (131)	3635 (176)	3410 (273)	3417 (262)	2132 (194)	3510 (287)	3728 (216)	3587 (277)	3241 (216)
	#	19	18	20	17	19	19	16	17	19	14
	*F*0	265 (44)	253 (38)	264 (40)	260 (39)	258 (37)	260 (43)	268 (42)	266 (42)	281 (61)	281 (40)
7	*F*1	984 (70)	882 (78)	815 (94)	812 (76)	794 (79)	634 (59)	579 (47)	425 (46)	624 (66)	449 (82)
	*F*2	1536 (90)	2441 (67)	1642 (148)	1223 (145)	2412 (90)	1637 (130)	2602 (111)	3002 (191)	1542 (136)	1700 (236)
	*F*3	3137 (378)	3344 (147)	3515 (128)	3354 (235)	3440 (214)	2253 (195)	3533 (192)	3618 (219)	3499 (224)	3205 (151)
	#	41	39	39	40	36	38	40	40	38	39
	*F*0	247 (32)	242 (35)	248 (38)	241 (30)	249 (30)	245 (31)	251 (29)	257 (32)	270 (40)	265 (39)
8	*F*1	969 (112)	873 (73)	740 (72)	766 (72)	782 (99)	600 (51)	560 (57)	414 (60)	607 (44)	458 (52)
	*F*2	1522 (186)	2370 (133)	1553 (162)	1181 (92)	2328 (118)	1695 (147)	2568 (130)	3031 (187)	1432 (128)	1577 (271)
	*F*3	3188 (312)	3254 (199)	3421 (209)	3392 (307)	3409 (282)	2212 (187)	3420 (321)	3626 (285)	3331 (216)	3165 (261)
	#	37	42	39	36	36	33	33	37	34	34
	*F*0	255 (37)	246 (32)	253 (40)	251 (31)	253 (33)	256 (41)	264 (39)	262 (45)	279 (46)	272 (45)
9	*F*1	1011 (77)	872 (75)	793 (75)	756 (76)	797 (77)	618 (55)	583 (58)	382 (62)	626 (54)	471 (73)
	*F*2	1601 (126)	2319 (143)	1529 (137)	1170 (98)	2287 (93)	1577 (142)	2522 (133)	2979 (147)	1472 (124)	1603 (225)
	*F*3	3245 (234)	3196 (177)	3391 (222)	3375 (233)	3390 (201)	2104 (205)	3466 (210)	3536 (252)	3340 (189)	3198 (224)
	#	42	41	38	39	37	40	39	43	41	37
	*F*0	257 (38)	243 (34)	251 (33)	241 (34)	250 (40)	242 (33)	253 (39)	254 (33)	273 (42)	268 (37)
10	*F*1	970 (89)	904 (78)	751 (66)	748 (51)	758 (69)	621 (43)	590 (53)	424 (69)	631 (59)	482 (63)
	*F*2	1558 (189)	2269 (103)	1628 (147)	1162 (101)	2254 (139)	1644 (149)	2401 (132)	2959 (167)	1501 (174)	1656 (268)
	*F*3	3148 (202)	3214 (203)	3318 (205)	3198 (318)	3407 (190)	2170 (211)	3434 (195)	3475 (151)	3267 (173)	3049 (314)
	#	38	38	39	41	39	40	37	42	37	37
	*F*0	250 (32)	244 (32)	243 (34)	241 (36)	244 (35)	239 (33)	254 (37)	250 (33)	267 (41)	265 (38)
11	*F*1	900 (76)	875 (67)	741 (55)	742 (54)	725 (58)	574 (54)	540 (37)	400 (60)	589 (37)	475 (53)
	*F*2	1436 (146)	2170 (88)	1525 (129)	1086 (76)	2136 (98)	1539 (121)	2406 (93)	2894 (133)	1479 (179)	1704 (232)
	*F*3	3075 (257)	3108 (255)	3166 (250)	3182 (275)	3176 (242)	2013 (186)	3290 (219)	3437 (236)	3072 (168)	2939 (180)
	#	40	37	41	39	37	38	32	36	38	34
	*F*0	233 (32)	221 (34)	225 (33)	220 (32)	225 (32)	225 (34)	228 (34)	232 (35)	243 (31)	239 (37)
12	*F*1	891 (76)	818 (70)	718 (63)	716 (55)	714 (66)	574 (66)	517 (36)	358 (47)	581 (54)	424 (51)
	*F*2	1432 (146)	2090 (137)	1484 (149)	1083 (88)	2026 (188)	1530 (124)	2207 (152)	2755 (155)	1450 (157)	1576 (223)
	*F*3	2930 (207)	3089 (191)	3081 (161)	3039 (233)	3161 (237)	2007 (170)	3191 (181)	3349 (238)	3006 (200)	2805 (208)
	#	26	25	25	28	29	23	22	25	26	23
	*F*0	181 (52)	182 (49)	187 (51)	180 (50)	184 (48)	184 (50)	188 (50)	189 (53)	191 (54)	190 (52)
13	*F*1	793 (58)	734 (46)	660 (39)	662 (43)	657 (46)	545 (29)	512 (27)	360 (57)	542 (24)	420 (41)
	*F*2	1324 (120)	1959 (116)	1388 (164)	1032 (72)	1923 (135)	1392 (122)	2067 (100)	2523 (159)	1396 (120)	1418 (238)
	*F*3	2664 (202)	2825 (189)	2807 (240)	2831 (261)	2916 (269)	1925 (167)	2860 (174)	3212 (412)	2834 (267)	2672 (285)
	#	20	20	19	19	18	15	20	21	17	16
	*F*0	176 (45)	166 (43)	169 (48)	167 (47)	171 (46)	170 (46)	176 (45)	177 (42)	182 (50)	179 (45)
14	*F*1	844 (88)	767 (83)	665 (61)	679 (61)	673 (56)	508 (45)	513 (34)	350 (32)	559 (35)	401 (23)
	*F*2	1379 (122)	1982 (177)	1376 (71)	1052 (75)	1955 (157)	1484 (118)	2188 (190)	2671 (220)	1379 (86)	1537 (255)
	*F*3	2679 (163)	2792 (154)	2882 (199)	2829 (184)	2970 (228)	1874 (133)	3040 (222)	3340 (342)	2794 (187)	2645 (154)
	#	19	18	18	18	15	16	16	20	15	17
	*F*0	125 (36)	122 (30)	130 (36)	124 (34)	116 (11)	126 (35)	131 (46)	128 (32)	139 (37)	130 (37)
15	*F*1	731 (57)	676 (56)	600 (51)	617 (40)	609 (38)	499 (18)	478 (37)	310 (33)	519 (34)	343 (44)
	*F*2	1316 (68)	1728 (91)	1385 (134)	976 (47)	1720 (53)	1337 (65)	1992 (152)	2350 (123)	1335 (67)	1316 (209)
	*F*3	2507 (139)	2573 (111)	2657 (132)	2634 (246)	2648 (105)	1755 (133)	2757 (142)	2964 (268)	2556 (160)	2433 (134)

TABLE II. (*Continued*.)

Age		aa	ae	ah	ao	eh	er	ih	iy	uh	uw
	#	17	19	18	17	19	14	16	18	16	17
	$F0$	126 (30)	120 (24)	122 (26)	118 (20)	122 (24)	123 (21)	131 (26)	126 (25)	134 (29)	131 (25)
16	$F1$	741 (58)	684 (60)	596 (25)	600 (35)	599 (26)	472 (52)	451 (29)	296 (20)	497 (31)	348 (40)
	$F2$	1261 (78)	1762 (65)	1254 (108)	935 (56)	1766 (105)	1354 (80)	1944 (152)	2334 (177)	1296 (81)	1368 (231)
	$F3$	2627 (209)	2620 (98)	2682 (185)	2737 (240)	2701 (90)	1754 (88)	2735 (140)	3030 (245)	2611 (114)	2397 (237)
	#	17	16	16	16	16	10	15	17	15	12
	$F0$	129 (26)	122 (21)	126 (23)	126 (22)	128 (25)	125 (21)	133 (28)	131 (23)	143 (31)	132 (23)
17	$F1$	713 (43)	685 (63)	585 (20)	612 (32)	581 (25)	491 (38)	457 (28)	289 (26)	514 (28)	322 (17)
	$F2$	1221 (126)	1759 (83)	1341 (136)	940 (27)	1745 (112)	1284 (42)	1910 (125)	2268 (110)	1348 (100)	1216 (281)
	$F3$	2637 (127)	2541 (77)	2600 (89)	2689 (187)	2651 (122)	1704 (89)	2724 (154)	3092 (212)	2573 (140)	2468 (242)
	#	15	17	16	17	16	13	16	17	14	8
	$F0$	124 (27)	123 (29)	122 (28)	125 (26)	120 (27)	123 (27)	136 (29)	123 (23)	142 (42)	134 (32)
18	$F1$	737 (48)	686 (49)	602 (26)	599 (40)	604 (40)	490 (25)	449 (35)	283 (20)	533 (32)	337 (19)
	$F2$	1269 (61)	1759 (93)	1252 (70)	881 (52)	1776 (69)	1282 (61)	1955 (117)	2289 (118)	1297 (75)	1144 (169)
	$F3$	2560 (160)	2560 (96)	2673 (182)	2622 (222)	2633 (89)	1625 (139)	2670 (108)	3050 (238)	2535 (63)	2328 (124)
	#	47	47	44	43	46	30	39	49	43	32
	$F0$	135 (31)	123 (27)	130 (30)	127 (30)	129 (27)	134 (30)	136 (32)	132 (28)	149 (31)	144 (33)
19+	$F1$	723 (48)	669 (43)	610 (32)	601 (33)	590 (32)	471 (29)	458 (26)	292 (26)	501 (35)	342 (34)
	$F2$	1204 (68)	1725 (100)	1288 (119)	929 (62)	1707 (114)	1265 (74)	1851 (110)	2266 (139)	1269 (121)	1185 (117)
	$F3$	2496 (176)	2532 (151)	2557 (156)	2599 (165)	2549 (134)	1612 (108)	2588 (105)	2930 (184)	2466 (156)	2411 (160)

that the within-subject pitch variation significantly decreases during two periods: from age 6 to age 8 and from age 12 to age 15 (i.e., during puberty). For female speakers, a significant difference in within-subject variability exists between age 10 and age 14 ($t=2.95$, $df=22$, $p<0.01$).

After puberty, the average within-subject $F0$ variation tends to increase again with age for both male and female speakers. Therefore, overshoot of acoustic parameters before reaching adult target levels may occur not only in duration but also in $F0$ variability. An interesting finding is that for subjects older than age 14, the within-subject $F0$ variability is significantly higher in female speakers than in male speakers [$F(1,138)=37.3$, $p<0.001$].

The mean $F0$ of individual vowels is shown in Fig. 3(d) for male speakers. It can be seen that male children as young as 5 years exhibit adult-like vowel-dependent $F0$ patterns. Similar observations can be made for female children (not shown here). This suggests that the capability of intrinsic $F0$ control in a given context is acquired earlier than age 5.[3]

C. Formant frequencies

Mean formant frequency values for the first three formants, averaged across all subjects for each vowel and age group, are provided in Tables II and III for male and female speakers, respectively. Two-sigma ellipses for five vowels are shown in Fig. 4(a) and (b) in the $F1-F2$ space for male and female speakers ages 10 to 12, and compared with results from similar studies in the literature. It is observed that the vowel positions produced by children of ages 10 through 12 in the current database are slightly compressed or centralized, compared to children's formant data in Peterson and Barney[4] (1952). This centralization of vowel space is most possibly due to the context difference (i.e., /hVd/ vs /bVt/) as well as dialect differences between the speaker population of the two studies (i.e., midwestern vs Pacific eastern). The most noticeable difference in formant distributions between the current study and the study by Hillenbrand *et al.*[5] (1995) is the proximity of front vowels /IY/ and /AE/ in the latter. It is also observed that despite the differences in vowel-class centroid values in the $F1-F2$ space, within-vowel variances (i.e., orientation of ellipses) are consistent between these studies and, clearly, the vowel $F1-F2$ space is larger for children than for adults.

Scatter plots of mean $F1$ and $F2$ of several vowels are shown in Fig. 4(c) for male speakers and in Fig. 4(d) for female speakers. Each point represents $F1$ and $F2$ values averaged across all subjects in the 5–6, 7–8,..., 17–18 age groups and adults. For instance, the rightmost circle in /IY/ represents mean $F1$ and $F2$ for children of ages 5 and 6, and the leftmost circle represents adults. A linear-scaling trend of male formant frequencies as a function of age is clearly observable from Fig. 4(c), especially between age 11 and age 16. Therefore, the current formant data of male speakers seem to support the notion of ''uniform axial growth of the vocal tract,'' as discussed in Kent (1976). Such a linear trend is, however, not clear for female speakers [Fig. 4(d)]. These trends can be more clearly observed in terms of the formant-scaling factors shown in Fig. 6.

Formant variability in terms of the within-subject COV is shown in Fig. 5(a) for male speakers and Fig. 5(b) for female speakers. The age-dependent reduction of the normalized formant frequency variability agrees with the trend shown in Eguchi and Hirsh (1969). Formant frequency variability may reach adult level around age 14 simultaneously for all formants. It is also observed that the COV is different for $F1$, $F2$, and $F3$ for subjects younger than age 12; specifically, the variability in $F1$ is greater than the variability in $F2$ and $F3$. For male speakers, the COV of $F1$, $F2$, and $F3$ take similar values after age 12, while such a trend is not observed for female speakers.

TABLE III. Mean and standard deviation (in parentheses) of fundamental frequency and formant-frequency values for female speakers (# is the number of tokens).

Age		aa	ae	ah	ao	eh	er	ih	iy	uh	uw
	#	14	16	15	19	16	20	11	13	12	17
	*F*0	272 (37)	263 (32)	277 (29)	273 (36)	268 (31)	258 (34)	279 (38)	265 (44)	300 (32)	286 (40)
5	*F*1	1224 (64)	1055 (105)	956 (105)	921 (90)	894 (83)	687 (62)	685 (66)	466 (30)	698 (31)	501 (54)
	*F*2	1842 (141)	2613 (138)	1772 (108)	1337 (88)	2555 (130)	1707 (137)	2816 (170)	3019 (180)	1376 (65)	1709 (263)
	*F*3	3435 (387)	3348 (230)	3274 (395)	3675 (250)	3227 (342)	2350 (215)	3526 (289)	3644 (98)	3496 (324)	3332 (161)
	#	25	24	22	25	19	25	23	20	22	20
	*F*0	265 (53)	248 (43)	258 (38)	259 (41)	259 (44)	252 (40)	264 (39)	271 (45)	282 (41)	280 (58)
6	*F*1	1163 (53)	972 (121)	839 (52)	844 (76)	828 (58)	654 (52)	643 (66)	433 (71)	670 (61)	512 (67)
	*F*2	1771 (189)	2671 (183)	1754 (165)	1258 (100)	2661 (115)	1815 (124)	2791 (123)	2986 (151)	1590 (183)	1793 (370)
	*F*3	3447 (292)	3438 (317)	3759 (233)	3729 (322)	3421 (349)	2458 (381)	3441 (347)	3596 (128)	3665 (381)	3380 (176)
	#	36	28	35	38	21	32	28	35	34	30
	*F*0	281 (43)	273 (41)	278 (31)	265 (37)	275 (41)	269 (40)	278 (31)	276 (36)	298 (40)	281 (42)
7	*F*1	1067 (102)	1023 (87)	846 (70)	853 (88)	856 (46)	686 (55)	608 (44)	467 (70)	661 (44)	506 (70)
	*F*2	1647 (216)	2433 (143)	1723 (162)	1279 (131)	2428 (113)	1740 (184)	2642 (178)	3026 (181)	1566 (110)	1840 (255)
	*F*3	3493 (361)	3410 (278)	3425 (502)	3569 (336)	3458 (234)	2490 (301)	3482 (283)	3573 (175)	3327 (468)	3316 (255)
	#	14	13	14	16	16	19	11	10	12	14
	*F*0	273 (31)	268 (40)	268 (37)	261 (33)	264 (31)	260 (39)	280 (35)	274 (42)	303 (50)	294 (49)
8	*F*1	1108 (64)	1021 (84)	848 (60)	870 (84)	851 (76)	612 (69)	568 (50)	428 (46)	664 (77)	426 (80)
	*F*2	1660 (57)	2419 (154)	1693 (109)	1274 (73)	2363 (199)	1713 (235)	2674 (192)	2997 (201)	1450 (90)	1539 (165)
	*F*3	3144 (342)	3271 (282)	3534 (193)	3384 (226)	3263 (352)	2381 (378)	3552 (166)	3604 (164)	3560 (108)	3298 (217)
	#	38	41	38	38	37	37	28	35	38	37
	*F*0	267 (35)	248 (25)	267 (33)	252 (30)	260 (31)	253 (29)	265 (34)	264 (29)	292 (35)	278 (36)
9	*F*1	1063 (86)	948 (89)	801 (48)	810 (64)	818 (68)	643 (60)	587 (54)	455 (61)	652 (51)	505 (44)
	*F*2	1676 (199)	2415 (131)	1658 (144)	1250 (93)	2363 (133)	1757 (163)	2559 (112)	3061 (140)	1506 (107)	1764 (220)
	*F*3	3284 (275)	3330 (180)	3505 (189)	3387 (248)	3431 (230)	2298 (182)	3516 (205)	3626 (217)	3412 (262)	3247 (195)
	#	25	21	24	24	17	24	22	23	23	22
	*F*0	263 (41)	255 (48)	257 (38)	258 (44)	259 (45)	258 (42)	267 (45)	261 (35)	275 (50)	273 (44)
10	*F*1	1037 (95)	970 (68)	791 (79)	822 (92)	860 (62)	612 (72)	609 (51)	472 (45)	636 (51)	496 (54)
	*F*2	1663 (251)	2318 (140)	1748 (118)	1264 (112)	2306 (92)	1733 (135)	2491 (178)	2969 (134)	1689 (240)	1747 (280)
	*F*3	3204 (272)	3286 (285)	3431 (227)	3378 (258)	3372 (312)	2264 (258)	3469 (334)	3506 (165)	3307 (270)	3166 (259)
	#	33	34	33	33	31	31	30	28	32	33
	*F*0	247 (37)	242 (40)	246 (38)	243 (37)	242 (37)	247 (41)	254 (39)	255 (37)	279 (47)	254 (34)
11	*F*1	980 (83)	878 (84)	765 (58)	791 (59)	775 (50)	638 (50)	590 (49)	467 (57)	637 (39)	475 (38)
	*F*2	1547 (163)	2219 (138)	1676 (124)	1219 (66)	2245 (109)	1645 (116)	2468 (113)	2971 (96)	1540 (152)	1774 (200)
	*F*3	3130 (208)	3190 (149)	3282 (165)	3305 (240)	3365 (175)	2167 (148)	3377 (195)	3462 (156)	3207 (139)	3033 (128)
	#	39	35	38	36	35	34	36	37	35	36
	*F*0	234 (27)	226 (26)	230 (24)	229 (22)	228 (25)	228 (25)	237 (29)	234 (28)	253 (38)	242 (25)
12	*F*1	939 (108)	836 (128)	742 (100)	761 (56)	712 (87)	620 (47)	537 (73)	439 (52)	591 (58)	452 (42)
	*F*2	1612 (164)	2215 (188)	1679 (142)	1212 (120)	2189 (170)	1662 (167)	2398 (222)	2884 (182)	1593 (147)	1661 (309)
	*F*3	3077 (267)	3163 (261)	3287 (221)	3148 (235)	3287 (168)	2247 (398)	3273 (270)	3376 (256)	3148 (233)	2963 (193)
	#	22	23	21	23	21	20	21	22	21	20
	*F*0	251 (43)	238 (32)	245 (31)	238 (27)	244 (28)	237 (28)	252 (37)	247 (29)	264 (45)	255 (32)
13	*F*1	959 (71)	824 (60)	779 (62)	753 (49)	770 (49)	627 (37)	592 (42)	426 (62)	654 (32)	490 (51)
	*F*2	1552 (150)	2187 (134)	1715 (150)	1212 (106)	2199 (96)	1668 (141)	2382 (94)	2861 (77)	1557 (162)	1858 (198)
	*F*3	3072 (198)	3110 (163)	3245 (148)	3132 (215)	3238 (109)	2161 (187)	3296 (117)	3391 (136)	3146 (179)	2968 (117)
	#	19	19	17	18	18	17	16	18	17	17
	*F*0	225 (27)	217 (21)	221 (19)	219 (21)	221 (26)	224 (19)	231 (16)	226 (19)	248 (33)	232 (24)
14	*F*1	893 (88)	824 (77)	756 (55)	746 (58)	736 (50)	627 (57)	573 (59)	415 (28)	630 (44)	433 (35)
	*F*2	1556 (72)	2010 (100)	1619 (101)	1192 (116)	2032 (114)	1639 (91)	2189 (92)	2693 (130)	1595 (112)	1693 (186)
	*F*3	2904 (232)	2935 (175)	3019 (173)	2972 (292)	3103 (190)	2115 (181)	3075 (181)	3222 (190)	2966 (194)	2724 (152)
	#	18	19	18	19	18	18	16	19	18	17
	*F*0	220 (27)	218 (27)	225 (31)	220 (29)	219 (28)	222 (27)	226 (25)	229 (29)	246 (35)	238 (33)
15	*F*1	900 (48)	851 (86)	738 (42)	767 (39)	744 (36)	594 (38)	564 (51)	378 (27)	620 (53)	434 (22)
	*F*2	1541 (91)	1942 (99)	1646 (131)	1177 (74)	1971 (118)	1586 (112)	2171 (75)	2653 (153)	1533 (116)	1727 (254)
	*F*3	2753 (197)	2907 (150)	2931 (157)	2991 (236)	2998 (165)	1949 (109)	3024 (105)	3237 (183)	2827 (149)	2674 (107)
	#	18	20	20	20	21	19	18	19	20	18
	*F*0	225 (20)	222 (20)	231 (24)	217 (22)	226 (20)	227 (23)	234 (21)	233 (28)	255 (32)	240 (25)
16	*F*1	851 (55)	835 (68)	737 (44)	749 (43)	735 (47)	602 (43)	541 (46)	423 (52)	613 (48)	447 (35)
	*F*2	1412 (98)	2050 (100)	1620 (199)	1159 (88)	2003 (143)	1628 (120)	2207 (89)	2776 (147)	1617 (162)	1691 (232)
	*F*3	2896 (375)	3004 (144)	2991 (146)	2966 (195)	3042 (142)	2056 (126)	3072 (111)	3241 (135)	2927 (155)	2866 (275)

TABLE III. (*Continued*.)

Age		aa	ae	ah	ao	eh	er	ih	iy	uh	uw
	#	17	17	16	15	13	14	16	15	15	13
	*F*0	219 (20)	205 (20)	212 (24)	204 (21)	206 (15)	207 (17)	217 (21)	217 (17)	236 (26)	225 (22)
17	*F*1	922 (82)	845 (72)	735 (37)	751 (41)	728 (42)	552 (45)	558 (58)	390 (63)	623 (28)	424 (22)
	*F*2	1467 (160)	2007 (106)	1514 (171)	1166 (93)	2010 (73)	1581 (135)	2168 (84)	2717 (84)	1575 (198)	1715 (392)
	*F*3	2803 (135)	2867 (256)	2890 (160)	2921 (161)	2961 (144)	1995 (123)	3024 (165)	3290 (130)	2801 (146)	2685 (138)
	#	18	17	18	17	19	18	16	16	17	17
	*F*0	242 (15)	233 (16)	238 (14)	230 (16)	232 (18)	237 (13)	250 (17)	246 (16)	262 (24)	256 (19)
18	*F*1	932 (47)	914 (45)	741 (33)	777 (45)	754 (58)	619 (53)	587 (52)	418 (37)	605 (35)	480 (34)
	*F*2	1473 (144)	1955 (110)	1631 (123)	1165 (92)	2014 (117)	1569 (142)	2222 (33)	2801 (46)	1579 (140)	1771 (324)
	*F*3	2914 (153)	2946 (127)	3027 (104)	3042 (207)	3047 (112)	2058 (134)	3080 (117)	3305 (59)	2924 (138)	2860 (89)
	#	48	46	47	44	45	39	41	46	45	45
	*F*0	231 (40)	215 (36)	218 (35)	213 (29)	219 (35)	222 (34)	235 (40)	228 (30)	246 (40)	243 (36)
19+	*F*1	894 (76)	787 (66)	740 (56)	726 (47)	694 (52)	543 (43)	532 (59)	360 (45)	595 (62)	412 (48)
	*F*2	1459 (124)	2078 (102)	1609 (135)	1079 (89)	2057 (123)	1481 (132)	2183 (111)	2757 (145)	1522 (140)	1388 (248)
	*F*3	2950 (252)	2916 (145)	2957 (161)	2986 (220)	3005 (139)	1884 (144)	3064 (136)	3291 (200)	2887 (155)	2804 (235)

In Fig. 6, formant-scaling factors, as defined in Fant (1975), computed from formant frequencies averaged across vowels are plotted as a function of age for male and female speakers. It is observed that differentiation of male and female $F2$ and $F3$ patterns begins at around age 11 and the formants become fully distinguishable around age 15, i.e., after puberty in male speakers. Between ages 10 and 15, formant frequencies of male speakers decrease faster with age and reach much lower absolute values than those of female speakers. This suggests that the total growth and rate of growth of the vocal tract is greater in male speakers. On average, it is clear that formant values reach adult range

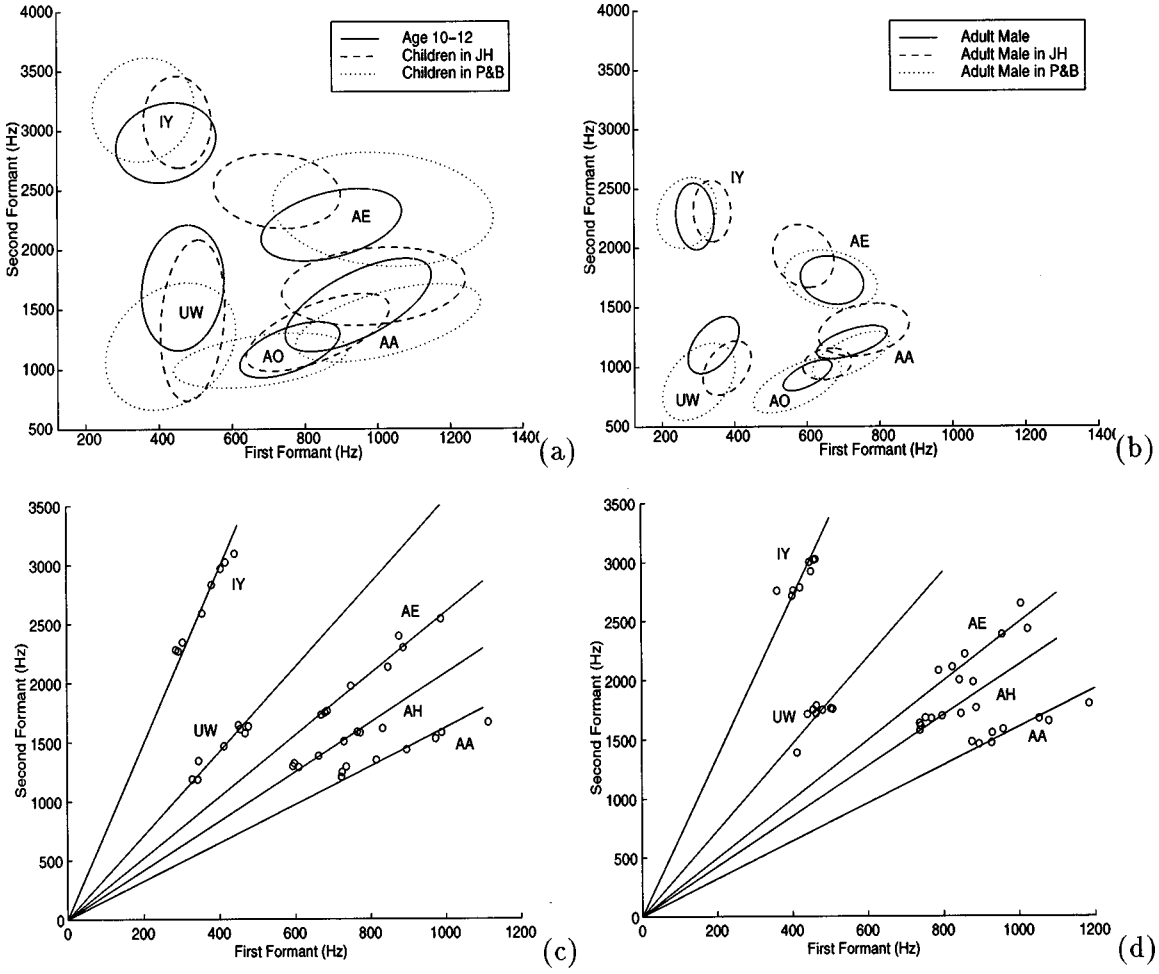

FIG. 4. Current formant data are compared with those in Hillenbrand *et al.* (1995) and Peterson and Barney (1959) for children (ages 10 through 12) and adults in (a) and (b), respectively. Plot of mean $F1$ and $F2$ of vowels /IY/, /AE/, /AA/, /AO/, and /UW/ are shown for males in (c) and for females in (d).

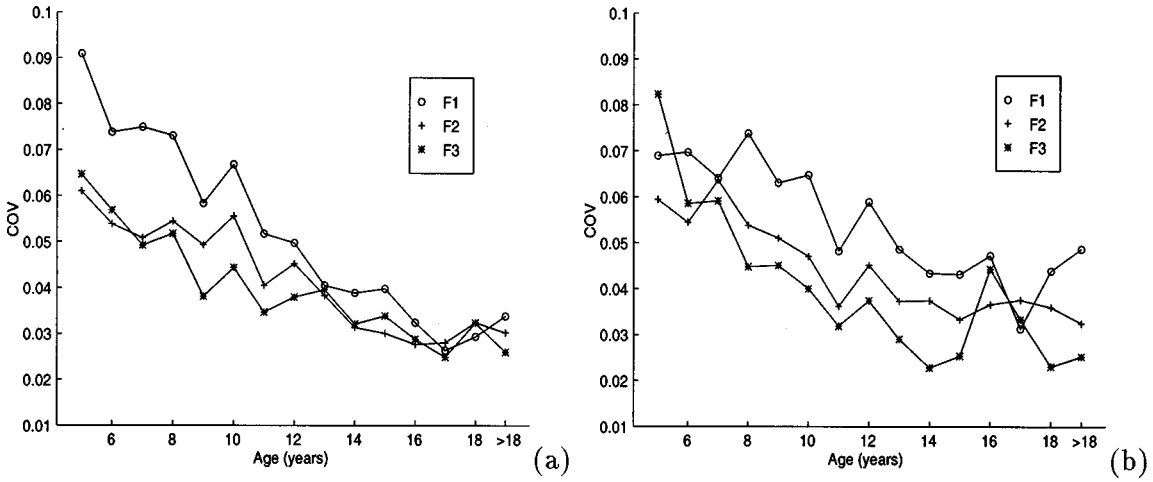

FIG. 5. Plot of within-subject COV of $F1$, $F2$, and $F3$ for (a) Male speakers and (b) Female speakers. It is clear that the within-subject formant variability decreases with age.

around age 15 for males and around age 14 for females. The formant scaling factors also suggest that the growth of vocal tract in male speakers may cease around age 15.

Formant scaling factors of male speakers are approximately the same for all formants and decrease almost linearly between ages 10 and 15. For female speakers, however, each formant evolves differently as a function of age. Differences in the rate of growth of the front and back cavities between female and male speakers could be a reason for this trend (Fant, 1975). Investigation of vowel-specific formant-scaling factors should help to verify this hypothesis.

D. Spectral-envelope variability

In Fig. 7(a), averaged cepstrum distances between two repetitions of the same vowel are shown as a function of age and gender. Vertical bars denote standard errors. The effect of age is significant [$F(14,476)=21.0$, $p<0.001$]. It is observed that spectral variability between vowel productions progressively decreases with age and converges to adult values around age 14 for both genders. Multiple comparisons indicate that variability is significantly reduced from age 5 to age 9 and from age 10 to age 14. There is no significant difference in variability after age 11. These results suggest that children younger that 10 years have not fully established their optimal or stable articulatory targets for vowels. The trend is similar to that of the formant variability shown in the previous section.

In Fig. 7(b), cepstrum distances between the initial and the final half of each vocalic segment are shown as a function of age for both male and female speakers. The effect of age is significant [$F(14,476)=9.8$, $p<0.001$]. Clearly, the within-vowel spectral variability progressively decreases with age. Multiple comparisons indicate that variability decreases significantly from age 10 to age 15. On average, children younger than age 10 display greater within-vowel spectral variability than adults.

As can be seen from Fig. 7(b), speakers display the least within-vowel spectral variation around age 15. Differences in spectral variations between age 15 and adults is significant ($t=2.39$, $df=38$, $p<0.03$). This suggests that in terms of the dynamics of /CVC/ articulations, an extremum in the production patterns may be achieved around age 15. It is also interesting to observe that female adults show significantly larger within-vowel spectral variation than male adults ($t=2.28$, $df=54$, $p<0.03$).

IV. DISCUSSION

The results of this cross-sectional acoustic study of children's speech, collected from subjects with no known speech pathologies, confirm that the reduction of magnitude and within-subject variability of temporal and spectral parameters with age is a general acoustic phenomenon associated with speech development in children. Furthermore, due to the wider age range of subjects (ages 5–18 years) examined in the current study, it was possible to obtain detailed acoustic information regarding how acoustic properties of chil-

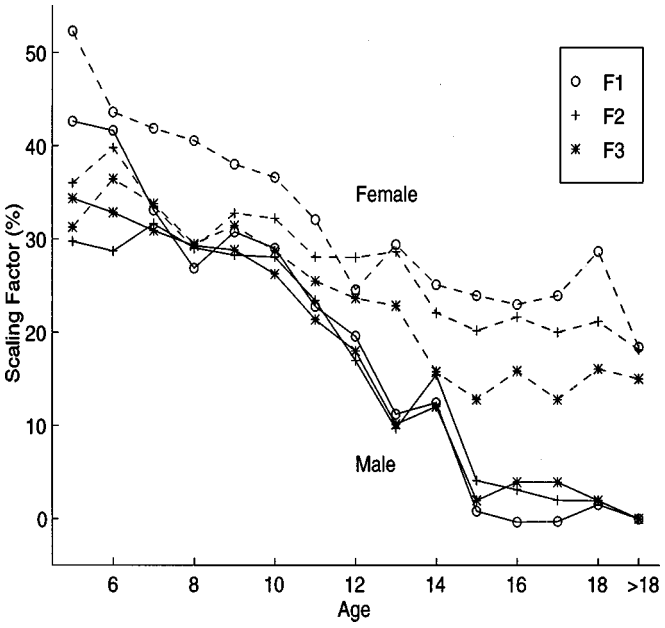

FIG. 6. Formant-scaling factor for $F1$ (○), $F2$ (+), and $F3$ (∗) scaled by adult male formants. They are computed based on mean formants averaged across all vowels and subjects in each age group.

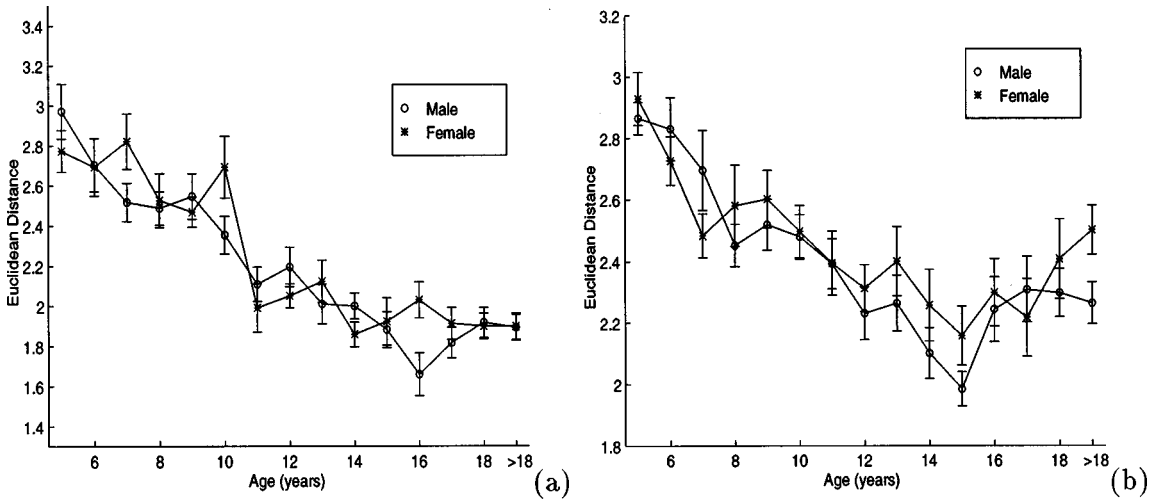

FIG. 7. (a) Mean cepstrum distance between the two repetitions of the same vowels is shown for each age group. (b) Mean cepstrum distance between the first- and second-half segments within vowel is shown for each age group.

dren's speech vary with age and when children begin to exhibit adult-like patterns. The acoustic data obtained in the current study are compared with similar acoustic data published in the literature, and the significance of the age- and gender-dependent acoustic trends observed in our data is discussed.

The main goal of this section is to explore the significance of the acoustic findings of this study based on (1) the physical development of the vocal tract and the voice source, and (2) the neuromuscular factors believed to govern the underlying articulatory dynamics, and hence the acoustic characteristics of speech. It should be noted that some of these interpretations remain to be validated and hence may be deemed speculative.

A. Comparison with previous studies

The acoustic data and age-dependent trends observed in this study are consistent with previously published acoustic data on children, in spite of differences in subject population and data-analysis procedures. For instance, trends in vowel durations measured in the current study and those in Hillenbrand *et al.* (1995) are similar, with an averaged correlation of 0.82 between the two studies for matched age groups. Hillenbrand *et al.* report durations for vowels embedded in /hVd/ target words produced in isolation for children of ages 10 through 12 and adults. Hillenbrand *et al.* also found that adult female speakers show significantly greater vowel durations than male adults. Although not statistically significant, the current data also show a similar trend. Finally, note that the within-subject variability of /s/ duration for male adults is comparable to that measured in Klatt (1972).

As illustrated in Fig. 4(a) and (b), the current formant data is in general agreement with those in Peterson and Barney (1952) as well as in Hillenbrand *et al.* (1995), except for the deviation of the centroids attributed to contextual and dialect differences. Regarding $F0$ measurements, it can be shown that the magnitude of $F0$ as well as the intrinsic $F0$ patterns are also quite similar to those in the two previous studies for matched age groups. Finally, the cross-sectional trend for the onset of pubertal $F0$ change is in agreement with the longitudinal study of 65 male speakers by Hollien *et al.* (1994), which reported that for the majority of subjects the pubertal-pitch change started between the ages of 12.5 and 14.5 years (mean 13.4) and was completed within 0.5 to 4.0 years (mean 1.5).

B. On *F*0 patterns

Average $F0$ and within-subject variability are highly correlated for speakers older than 8 [compare Fig. 3(a) and (c)]. It is thus possible that $F0$ variability can be fully predicted from $F0$ magnitude beyond a certain stage in the development process (e.g., age 8). Using the COV instead of the un-normalized variability measure does not change the essence of the observations made above. It is interesting to note, however, that from age 5 to 8 there is a 50% decrease in variability with essentially no change in average $F0$, suggesting a developmental stabilization of pitch control. Furthermore, female speakers over age 14 show significantly larger within-subject $F0$ variability than male speakers [Fig. 3(b)]. Although the exact reason for this is unclear, this trend may be due to inherent differences in the vocal-fold physiology (e.g., mass and length) of a high-$F0$ voice compared to a low-$F0$ voice.

C. On formant-scaling behavior

Growth curves for the larynx, pharynx, oral cavity, and total vocal-tract length are correlated with the average vowel fundamental frequency, average $F1$, $F2$, and $F3$ values, respectively (see Figs. 3 and 6). For example, $F3$ values agree with those predicted from Goldstein's (1980) model for vocal-tract-length growth in children (at least up to age 15). Good agreement has also been found between average age-dependent $F1$ values and growth of the back vocal-tract cavity (Fant, 1975). However, the limitations imposed by the physical dimensions of the vocal tract on the dynamic range of formants are much more stringent on $F2$ and $F3$ than $F1$. As a consequence, there is greater room for variation in average $F1$ values (averaged over all vowels) of the (child) speaker. In such cases, average $F1$ values may not always be

closely correlated with the physical size of the back oral cavity. For example, the difference in average $F1$ values between young male and female children, and the sudden drop of $F1$ from teenagers to adult females seen in Fig. 6 are not supported by physical growth data. We speculate that social and psychological factors contribute to these deviations in $F1$ values.

Overall, it can be deduced from Figs. 3 and 6 that physical growth of the speech apparatus occurs gradually up to approximately age 14 for females and age 15 for males. For male speakers, a growth spurt occurs somewhere between ages 12 and 15 (puberty) lasts about 1.5 years, and affects not only the larynx but the entire vocal-cavity size. In the developmental model of Goldstein (1980), however, in male speakers the vocal tract is assumed to continue growth beyond age 15 (14.6 cm) until age 20 (16.6 cm), a fact which is not supported by the current formant scaling factors. It does not seem plausible that the vocal-tract length could grow further beyond age 15 without inducing any decrease in the formant frequencies. Further investigation is needed to find the source of this discrepancy.

D. Significance of acoustic findings on underlying articulatory dynamics

First, consider the nearly universal trend of significant reduction in segmental durations, and their convergence to adult values, during the period from around age 9 or 10 to age 12 or 13. This is true irrespective of the segmental levels considered in this study, i.e., vocalic segment [Fig. 1(a)], fricative /s/ [Fig. 2(a)], and sentences [Fig. 2(c)]. Since systematic reduction of duration with age can be concomitant with improvements in speed of articulatory movement, the observation suggests that the underlying neuromuscular control of articulators also rapidly improves and converges to adult levels during that period. Accumulated experience in speech production could also play a role toward contributing to decreased segmental durations.

Second, consider the substantial reduction in within-subject duration variability between ages 8 to 14 years for vowels, the fricative /s/, and sentences [Figs. 1(b), 2(b), and (d)]. Along the lines of the discussion above, this finding may be interpreted as suggesting an *improved* articulatory-timing control being achieved between ages 8 and 14. It is thus thought that the progressive reduction in both duration magnitude and variability prior to age 12 is actually due to improvements in speech motor-timing control with age, and not merely an artifact of longer durations as argued in Kent and Forner (1980), and Crystal and House (1988). The poor correlation between the within-subject variability of /s/ and the mean sentence duration in the current data offers supporting evidence in this direction. For instance, the Pearson's correlation coefficient is 0.11 for adult speakers and 0.09 for children of ages 7 and 8, indicating almost no correlation between an individual's speaking rate and within-subject variability of /s/.

Third, consider the significant reduction in within-subject formant and spectral-envelope variability between about ages 7 and 11 as seen in Figs. 5 and 7(a), respectively. Since both formant structure and spectral-envelope shape are directly related to an underlying articulatory configuration, the reduced variability in these parameters may be attributed to reduced variability when reaching the individual's canonical articulatory configurations (targets) for a given sound. According to this interpretation of the formant and spectral-envelope data, articulatory ''robustness'' may be almost fully achieved by age 11 or 12. But, as discussed above, variability in reaching articulatory targets may also be correlated with speaking rate. Since the exact relation is not known, it is difficult to speculate further on the development of articulatory target-attainment reliability.

Finally, consider the trend wherein duration magnitude and variability (Figs. 1 and 2) reach minima somewhere between ages 13 and 15 before increasing and converging toward adult range. A similar trend is observed in within-subject formant and spectral-envelope variability (Figs. 5 and 7), which attain minima around 14 or 15 years of age. The smaller pitch variation for teenagers than for adults in Fig. 3(c) offers an additional attestation to this trend. If we assume the final (target) acoustic-parameter values at the completion of human speech development to be those of adults, these findings may be interpreted as demonstrating an apparent *overshoot* phenomenon of acoustic parameters before converging to their final values. Further, these findings may be interpreted as suggesting that teenagers of ages 14 and 15 exhibit an *extremum* in the production patterns among all age groups examined in this study, in terms of both speed of articulatory movements and (to a lesser extent) consistency in achieving desired articulatory configurations. Why that should be the case is not clear, however. This phenomenon may be associated with the learning process. Or it just may be that these patterns reflect the developmental nature of human physiological functions which underlie articulation and peak during the teenage years. Clearly, these overshoot phenomena have to be taken into account when interpreting age-dependent trends in acoustic data.

V. CONCLUDING REMARKS

Not all findings in our acoustic data could be accounted for or readily explained. Some of those cases follow: (1) One unexpected observation was the increase in between-subject variability in the sentence productions of speakers over 14 years old. This result may imply that the decrease in between-subject variability that occurs during speech development ceases after the acquisition of adult-like speech-production patterns simply due to the dominating effects of individual differences (or habit) in speaking rate (causing larger individual deviations from the group mean). (2) There is a significant decrease in $F1$ between the 18-year-old and adult female speakers. This trend was verified by manual estimation and comparison of formant frequencies from a subset of vowel tokens produced by the two age groups. This could be a sociolinguistic phenomenon. (3) There is a rebound and a drop in formant scaling factors between ages 12 and 14 for females, and between ages 13 and 15 for male speakers. This may be retrospective behavior on the part of speakers during puberty. The same trend has also been observed in the within-subject $F0$ variability, which could be the result of a similar retrospective behavior manifested in

$F0$. (4) $F1$ values for young female and male children are significantly different, far larger than any that could be explained by the physiological differences at this age.

Finally, it is noted that the uneven distribution of subjects by age and gender may cause a problem when comparing means of different age groups. This and the fact that subjects of ages 5 and 6 produced the target words for vowels in isolation are flaws of this study, attributed to the design of the speech database-collection procedure.

As demonstrated by the overshoot of acoustic parameters discussed earlier, the maturation process may confound the age-dependent values of other acoustic parameters as well. Furthermore, certain acoustic parameters are associated with more than one aspect of speech development and growth. As a result, additional measurements and different speech-elicitation scenarios are required to map in detail all aspects of speech development. For example, in order to understand if greater temporal variability is due to less precision or simply due to increased exploration (as a result of learning), speech can be elicited in a scenario where children are explicitly asked to minimize temporal variability, as in Smith and Kenney (1994). Direct articulatory and aerodynamic measurements could give clearer answers regarding development of motor control (Sharkey and Folkins, 1985; Smith and McLeane-Muse, 1986; Stathopoulos, 1995; Yang and Kasuya, 1994, 1995). Although many questions in speech development are yet to be answered, the current study nevertheless provides a better insight into the acoustic modeling of children's speech than was available before. From the acoustic modeling point of view, especially for developing speech applications such as automatic speech and speaker recognition, the increased variability in children's speech is a fact that one has to cope with independent of its underlying source.

ACKNOWLEDGMENTS

The authors wish to thank Andrej Ljolje at AT&T for his help with the automatic-segmentation program, Jay Wilpon at AT&T for his support throughout the course of this work, and Jim Miller and Rosalie Uchanski at CID for encouragement and helpful discussions. We also thank Anders Lofquist and Abigail Kaun for their clear and thorough comments which were essential in improving the form and content of this paper.

[1]The age distribution of the tokens discarded from the duration analysis was computed to investigate possible age- or gender-specific bias. Despite the fact that a higher percentage of tokens was discarded for children age 10 and below than for adults (25% versus 21%), no particular age group had more than 25% of the tokens discarded.

[2]Distribution of tokens discarded from formant analysis showed that a relatively small age-dependent trend exists: About 20% more tokens were discarded for children younger than 10 years than for adults.

[3]Because the tongue body and jaw positions affect the position of the larynx and the tension of the vocal folds (and thus $F0$ values), it is possible that the ability of intrinsic $F0$ control does not require any special explicit control mechanism, but is the result of the acquisition of proper articulatory positions for the vowels (cf. Whalen and Levitt, 1995).

[4]Peterson and Barney (1952) used a total of 15 child subjects in their analysis. Although the age range of the children is not provided in their paper, informal comparison with the formant ellipses of Peterson and Barney indicated the best match was with the 8-year-old children in our data.

[5]The study of Hillenbrand *et al.* (1995) was based on a total of 46 subjects ages 10 to 12 years.

Crystal, T. H., and House, A. S. (**1988**). ''A note on the variability of timing control,'' J. Speech Hear. Res. **31**, 497–502.

Davis, S., and Mermelstein, P. (**1980**). ''Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,'' IEEE Trans. Acoust., Speech, Signal Process. **28(4)**, 357–366.

Eguchi, S., and Hirsh, I. J. (**1969**). ''Development of speech sounds in children,'' Acta Oto-Laryngol. Suppl. **257**, 1–51.

Fant, G. (**1975**). ''Non-uniform vowel normalization,'' STL-QPSR 2-3/1975, 1–19.

Goldstein, U. G. (**1980**). ''An articulatory model for the vocal tracts of growing children,'' Ph.D. thesis (MIT, Cambridge, MA).

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.

Hollien, H., Green, R., and Massey, K. (**1994**). ''Longitudinal research on adolescent voice change in males,'' J. Acoust. Soc. Am. **96**, 2646–2654.

Kent, R. D. (**1976**). ''Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic study,'' J. Speech Hear. Res. **19**, 421–445.

Kent, R. D., and Forner, L. L. (**1980**). ''Speech segment durations in sentence recitations by children and adults,'' Journal of Phonetics **8**, 157–168.

Klatt, D. H. (**1974**). ''The duration of /s/ in English words,'' J. Speech Hear. Res. **17**, 51–63.

Ljolje, A., and Riley, M. D. (**1991**). ''Automatic segmentation and labeling of speech,'' *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (Toronto, Canada), pp. 473–476.

Miller, J. D., Lee, S., Uchanski, R. M., Heidbreder, A. H., Richman, B. B., and Tadlock, J. (**1996**). ''Creation of two children's speech databases,'' Proceedings of the ICASSP (Atlanta, GA), pp. 849–852.

Palethorpe, S., Wales, R., Clark, J. E., and Senserrick, T. (**1996**). ''Vowel classification in children,'' J. Acoust. Soc. Am. **100**, 3843–3851.

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

Potamianos, A., Narayanan, S., and Lee, S. (**1997**). ''Automatic speech recognition for children,'' *Proceedings of European Conference on Speech, Communication and Technology* (Rhodes, Greece), pp. 2371–2734.

Rabiner, L. R., and Juang, B-H. (**1993**). *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ).

Secrest, B. G., and Doddington, G. R. (**1983**). ''An integrated pitch tracking algorithm for speech systems,'' Proceedings of the ICASSP (Boston, MA), pp. 1352–1355.

Sharkey, S. G., and Folkins, J. H. (**1985**). ''Variability of lip and jaw movements in children and adult: Implications for the development of speech motor control,'' J. Speech Hear. Res. **28**, 8–15.

Smith, B. (**1978**). ''Temporal aspects of English speech production: A developmental perspective,'' Journal of Phonetics **6**, 37–67.

Smith, B. L. (**1992**). ''Relationships between duration and temporal variability in children's speech,'' J. Acoust. Soc. Am. **91**, 2165–2174.

Smith, B. L., and Kenney, M. K. (**1994**). ''Variability control in speech production tasks performed by adults and children,'' J. Acoust. Soc. Am. **96**, 699–705.

Smith, B., 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.

Smith, B. L., and McLeane-Muse, A. (**1986**). ''Articulatory movement characteristics of labial consonant productions by children and adults,'' J. Acoust. Soc. Am. **80**, 1321–1328.

Stathopoulos, E. T. (**1995**). ''Variability revisited: an acoustic, aerodynamic and respiratory kinematic comparison of children and adults during speech,'' Journal of Phonetics **23**, 67–80.

Whalen, D., and Levitt, A. (**1995**). ''The universality of intrinsic $F0$ of vowels,'' Journal of Phonetics **23**, 349–366.

Yang, C. S., and Kasuya, H. (**1994**). ''Accurate measurement of vocal tract shapes from magnetic resonance images of child, female, and male subjects,'' *Proceedings of the International Conference on Speech Language Processing* (Yokohama, Japan), pp. 623–626.

Yang, C.-S., and Kasuya, H. (**1995**). ''Uniform and non-uniform normalization of vocal tracts measured by MRI across male, female and child subjects,'' IEICE Trans. Inf. and Syst. **E78-D**, No 6, 732–737.