ANALYSIS OF CHILDREN'S SPEECH: DURATION, PITCH AND FORMANTS

Sungbok Lee, Alexandros Potamianos and Shrikanth Narayanan*

AT&T Labs–Research, 180 Park Ave, P.O. Box 971, Florham Park, NJ 07932-0971, U.S.A.
email: {sungbok,potam,shri}@research.att.com

ABSTRACT

Magnitude and variability of duration, pitch and formant frequencies are computed for speech collected from five to eighteen year-old children. The study confirmed that reduction in magnitude and variability are the primary indicators of speech development. Specifically, children below age ten exhibit wider dynamic range of vowel duration, longer suprasegmental duration, and larger temporal and spectral variations. These trends diminish around age twelve. Children's speech acoustic characteristics fully develop to adult level in both magnitude and variability around age fifteen. Change of formant frequencies in male speakers parallels the growth of the vocal tract, while for female speakers the presence of such a linear trend is not clear. We conclude that the primary factors governing the acoustic patterns during speech development are anatomical maturation of the speech apparatus and speech motor control in terms of agility and precision.

1. INTRODUCTION

An important issue in the study of speech development in children is to investigate the age at which magnitude and variability of acoustic parameters reach adult range and how these acoustic parameters vary as a function of age, gender, and speech sounds. In-depth knowledge on age–dependent acoustic patterns and their variability should be valuable for speech applications such as automatic recognition of children's speech [1], evaluation and training of deaf children, and text–to–speech synthesis. Such knowledge is also important for associating acoustic development in children with the underlying anatomical development of the vocal organs, which is essential for the creation of a better developmental model of the vocal tract for speech production research [2].

In this paper, speech duration, pitch and formant frequencies are measured together with temporal and spectral variability using a recently collected children's speech database [4]. The database enables a cross–sectional study from age five through age eighteen with an approximate one-year age resolution of one–year, filling the age gap that existed in previous studies (eg.,[3]) and providing a more detailed view of speech development. This paper is organized as follows: The database used in this study is briefly described in Section 2. In Section 3, the procedure used to estimate the acoustic parameters is described and evaluated. The results of this experimental study are presented in Section 4 followed by a discussion of speech development in children.

*Central Institute for the Deaf, St. Louis, MO 63110, U.S.A.

2. SPEECH DATABASE

The database used in this study was collected from 436 children of age five through age eighteen, with resolution of one–year of age, as well as 56 adults, of both genders. The speech material consisted of ten monophthongs and five diphthongs of American English vowels and five phonetically–balanced sentences repeated twice by all subjects. Target words for vowels were produced in the carrier sentence "I say uh – again," except for children of ages five and six. The target words for the monophthongal vowels are *bead (IY), bit (IH), bet (EH), bat (AE), pot (AA), ball (AO), but (AH), put (UH), boot (UW), and bird (ER)*. 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. A detailed description of subjects, data collection procedures, and the recording environment can be found in [4].

3. MEASUREMENTS

In this section, the automatic procedure used to measure the duration, pitch and formant frequencies is presented and evaluated by comparing its results to corresponding hand–measured data points for a small subset of the data. The automatic segmentation procedure used to obtain the phonemic, word and sentence boundaries is also outlined.

3.1. Duration

In order to process the large number of waveforms (over 23,000 files) within a reasonable amount of time, the phonetic level segmentation of each waveform was automatically computed by aligning the speech frames to the states of the corresponding phonemic units (hidden Markov models) trained from the children speech database [5]. Duration was measured directly from the start and end points of the target segment under consideration with a 10 msec resolution.

The accuracy of the automatic segmentation was evaluated from hand–measured durations of 160 randomly selected vowel segments. Mean difference between the automatically computed and the hand–measured values was −17.5 msec and the standard deviation was 37.0 msec. In general, vowel durations were systematically underestimated by the automatic procedure and thus vowel durations obtained in this study may be somewhat shorter than their actual values. However, the automatic segmentation of /s/ and the end–point detections of the five sentences were quite accurate.

3.2. Fundamental and formant frequencies

The fundamental frequency (F0) and the first three formants (F1, F2, F3) of the ten vowel segments were computed using the automatic pitch and formant–tracking program of the commercial software package ESPS by Entropic Research Laboratory Inc. The *median value* of each pitch and formant track was computed as the representative value of the track. The performance of the automatic program was evaluated using a subset of hand–measured data. Mean differences (standard deviation) between hand and automatically measured 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. In general, pitch and formants were underestimated by the automatic program.

Formant values that were clearly erroneous were discarded using the following procedure: the initial formant data were grouped by vowel, age and gender, and outliers in each group were removed using a two–sigma ellipse. Next, each data file was visually examined and whenever one of the F1-F3 values was judged to be too low or too high, the corresponding formant set was discarded. Despite our efforts to remove erroneous formant values some might have escaped the visual inspection process. It is quite possible that the refined formant data set included some underestimated F2 and F3 values, especially for children age eight and lower. For pitch estimates, the large standard deviation of F0 estimation error was due to a few occurrences of "pitch halving" by the automatic program. These erroneous pitch values were removed by a procedure similar to the one used for the refinement of the formant data. These refined pitch and formant data are presented in this study.

3.3. Spectral envelope

The smooth spectral envelope or equivalently the cepstrum coefficients derived from the spectral envelope are the most common set of features used in automatic speech recognition. Thus, it is important to measure the spectral variability of speech sounds as a function of speaker's age. In the current study, spectral variability between two repetitions of a target vowel, and between the first and second half segments of each vowel segment were measured. For this purpose, a given vowel segment was analyzed using a mel–frequency filterbank (spanning 100 to 6000 Hz) and the first 12 cepstral coefficients were computed (plus energy). The Euclidean distance between the two set of coefficients was used as a measure of spectral variability.

3.4. Variability measurements

Variability of the temporal and spectral parameters associated within each age group was measured both within and between speakers. Inter–subject variability was computed as the standard deviation of the average value of a given parameter across all subjects in an age group. Intra–subject variability was computed as the difference of the magnitude of a given parameter between two repetitions. Group intra–subject variability was defined as the average intra-subject variability of that group. Finally, the coefficient of variation (COV) was computed as the ratio of inter– or intra–subject variability to the corresponding mean value.

4. RESULTS AND DISCUSSION

In this section, the estimated duration, pitch and formant frequencies are presented, together with temporal and spectral variability measurements. Implications for the maturation of the speech production process for children speakers are discussed.

4.1. Duration

Mean durations of 10 monophthongs are shown in Fig. 1(a) for various age groups. ANOVA analysis shows no statistically significant gender difference in vowel duration. As can be seen in Fig. 1(a), five year-old children display longer vowel durations than older age groups. There is no statistically significant change in vowel duration after age seven. It is interesting to observe that although children of ages five to seven exhibit adult–like pattern of vowel–dependent duration, their relative timing control among vowels in a given context is not well yet established. They show a tendency to overshoot, or sometimes possibly undershoot, vowel duration. This may suggest that the dynamic range of vowel duration is larger for young children than for adults.

The intra–subject COV is shown in Fig. 1(b). Vowel duration variability reaches adult levels around age eleven or twelve, several years later than vowel duration magnitude. A recent study on vowel duration also reported a similar trend [6]. The latency of variability stabilization may be explained as the time required to adjust the wider dynamic range of vowel duration.

In contrast, the analysis of /s/ and sentence duration (not shown in the plots) indicates that both magnitude and variability reach adult levels at approximately the same time, around age eleven or twelve. Fluent productions of sentences and /s/ (in "I say") might require better coarticulation skills than vowel production in a given context. These results imply that both duration and variability in suprasegmental levels may be governed by a single but collective or emergent factor, the degree of coarticulatory skill. It is interesting to note, that on average, teenagers around age fifteen show shorter duration than adults for both /s/ and for sentence productions.

4.2. Pitch

The average pitch of male and female speakers averaged across all vowels is shown in Fig. 1(c) as a function of age (inter–subject variation shown with error bars). No statistically significant gender difference exists in pitch up to age twelve. For male speakers, there is a significant F0 drop from age eleven to age thirteen and there is no significant pitch change after age fifteen. This indicates that pubertal pitch change in male children starts between ages twelve and thirteen, and ends around age fifteen. About a one-octave pitch drop is observed during puberty. The relatively large inter–subject variability at ages thirteen and fourteen suggest that the onset-time of puberty is speaker–dependent. For female speakers, the pitch drop from age seven to age twelve is significant, indicating that the laryngeal growth in females ends around age twelve and thus pitch reaches adult levels after that age. On average, teenagers after puberty show lower pitch values than adults. Intra–subject COV indicates that pitch variability progressively decreases with increasing age and reaches adult levels around age twelve or thirteen for both genders. Teenagers show, on average, less variability than adults.

4.3. Formant frequencies and scaling factors

Two–sigma ellipses of several vowels are shown in Fig. 1(d) in the F1/F2 space. The vowel positions produced by 8-year old boys in the current database (Fig. 1(d)) are slightly compressed or centralized, compared to the children's formant data in [7], most possibly due to the context difference (/h/–vowel vs. /b/–vowel) as well as dialect differences between the speaker population of the two studies. Scatter plots of mean F1 and F2 of several vowels produced by male speakers are shown in Fig. 1(e). Each point represents mean F1 and F2 averaged across all subjects in an age group. For instance, the rightmost circle in /iy/ represents mean F1 and F2 for children of ages five and six, and the leftmost circle for adults. In Fig. 1(f), formant scaling factors scaled by the mean formants of adult male speakers averaged across all vowels are plotted as a function of age.

As shown in Fig. 1(f), differentiation of male and female formant patterns begin around age ten or eleven and formants become fully distinguishable around age fifteen. Between age ten and fifteen, vowel formant frequencies of male speakers decrease faster and reach much lower values than those of female speakers, implying that both the total growth and the rate of growth of the pharynx is greater for male speakers. On average, formant values reach adult range around age fifteen for males and around age fourteen for females.

A linear-scaling trend of male formant frequencies as a function of age is clearly observable from Fig. 1(e). This implies that for male speakers, the acoustic changes resulting from vocal tract growth are uniform across vowel formants independent of articulatory differences among vowels. The mean scaling factors of male speakers are approximately the same for F1, F2, F3 and decrease almost linearly as vocal tract grows between ages nine and fifteen. Such a linear trend is, however, not clear for female speakers. Each formant evolves differently as a function of age and vocal tract growth. Another characteristic of female formant patterns is the sudden drop of F1 from age 18 to adult. Physiological and/or socio–psychological factors could be the reason for this phenomenon.

Overall, F3 drops by about 32% from age five to age fifteen for male speakers, as can be seen in Fig. 1(f). This is comparable to the change of vocal tract length (33%) from five year old male children to adults (age 20). predicted in [2]. In [2], however, substantial (10%) vocal tract length growth is predicted to occur between ages fifteen and twenty, which is not supported by our pitch and formant data. No substantial formant frequency change can be seen in Fig. 1(f) over age fifteen.

4.4. Spectral variability

In Fig. 1(g), the mean cesptrum distance between two repetitions of the same vowel by the same speaker is shown. Clearly, young children exhibit more spectral variation than adults between two repetitions of the same utterance. A similar trend can be also observed in formant variability. This suggest that young children, especially below age ten, have not fully established their own optimal articulatory vowel targets in a given context.

In Fig. 1(h), the mean cepstrum distance between the first and the second half of each vowel segment is shown as a function of age and gender. It is clear that even within a vowel utterance, children younger than ten display greater spectral variability. This is most likely due to excessive and/or abrupt tongue movement during the transition from the vowel to the final consonant /d/, /t/ or /l/, which signifies that children are less-skilled in coarticulation. It was also observed that these excessive variations disappear around age eleven or twelve where variability reaches adult levels. Since both phonemic and sentence duration variability (see Sec. 4.1) reach adult levels around that same age, temporal and spectral variability seem to follow parallel paths and are a measure of the maturity of the motor control for efficient co-articulation. Therefore, adult–like coarticulation skills are achieved around age twelve.

As can be observed from Fig. 1(h), teenagers around age fifteen display less within–vowel spectral change than all other age groups. This could be possibly due to the faster speaking rate for teenagers (see Sec. 4.1). Further, female adults display greater within–vowel spectral variation during transitions than male adults.

5. CONCLUSIONS

In this study, we have measured the duration, pitch, and formant frequencies of speech collected from speakers ages five through eighteen with a resolution of one–year of age. Reduction in magnitude and within–subject variability over time are two major indicators of speech development. Specifically, when compared to adult, children below age ten exhibit wider dynamic range of vowel duration, longer segmental and suprasegmental durations, higher pitch and formant values, and larger within–subject variability. This trend diminishes around age twelve and, *in both magnitude and variability*, children's speech fully develops to adult levels around age fifteen for male speakers and age fourteen for female speakers. Change of formant patterns in male speakers parallels the vocal tract growth, while for females such a linear trend is not clear. Teenagers around age fifteen differ from both children and adults in that they speak faster, have lower pitch values and exhibit less temporal and spectral variability. We conclude that the primary factors governing the acoustic patterns during speech development are anatomical maturation of the speech organs and speech motor control in terms of agility and precision.

6. ACKNOWLEDGMENTS

The authors wish to thank Jay Wilpon at AT&T for his support throughout the course of this work, and Jim Miller and Rosalie Uchanski at CID for helpful discussions.

7. REFERENCES

[1] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Proc. EUROSPEECH' 97*.

[2] U. G. Goldstein, "An articulatory model for the vocal tracts of growing children," *Ph.D. Thesis*, MIT, 1980.

[3] S. Eguchi and I. J. Hirsh, "Development of speech sounds in children," in *Acta. Otolaryng.*, Suppl. vol. 257, 1969.

[4] J. D. Miller, S. Lee, R. M. Uchanski, A. F. Heidbreder, B. B. Richman and J. Tadlock, "Creation of two children's speech databases," in *Proc. ICASSP*, pp. 849–852, 1996.

[5] A. Ljolie and M. D. Riley, "Automatic segmentation and labeling of speech," in *Proc. ICASSP*, pp. 473–476, 1991.

[6] B. L. Smith, "Relationships between duration and temporal variability in children's speech," in *JASA*, vol. 91, pp. 2165–2174, 1992.

[7] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," in *JASA*, vol. 24, pp. 175–184, 1952.

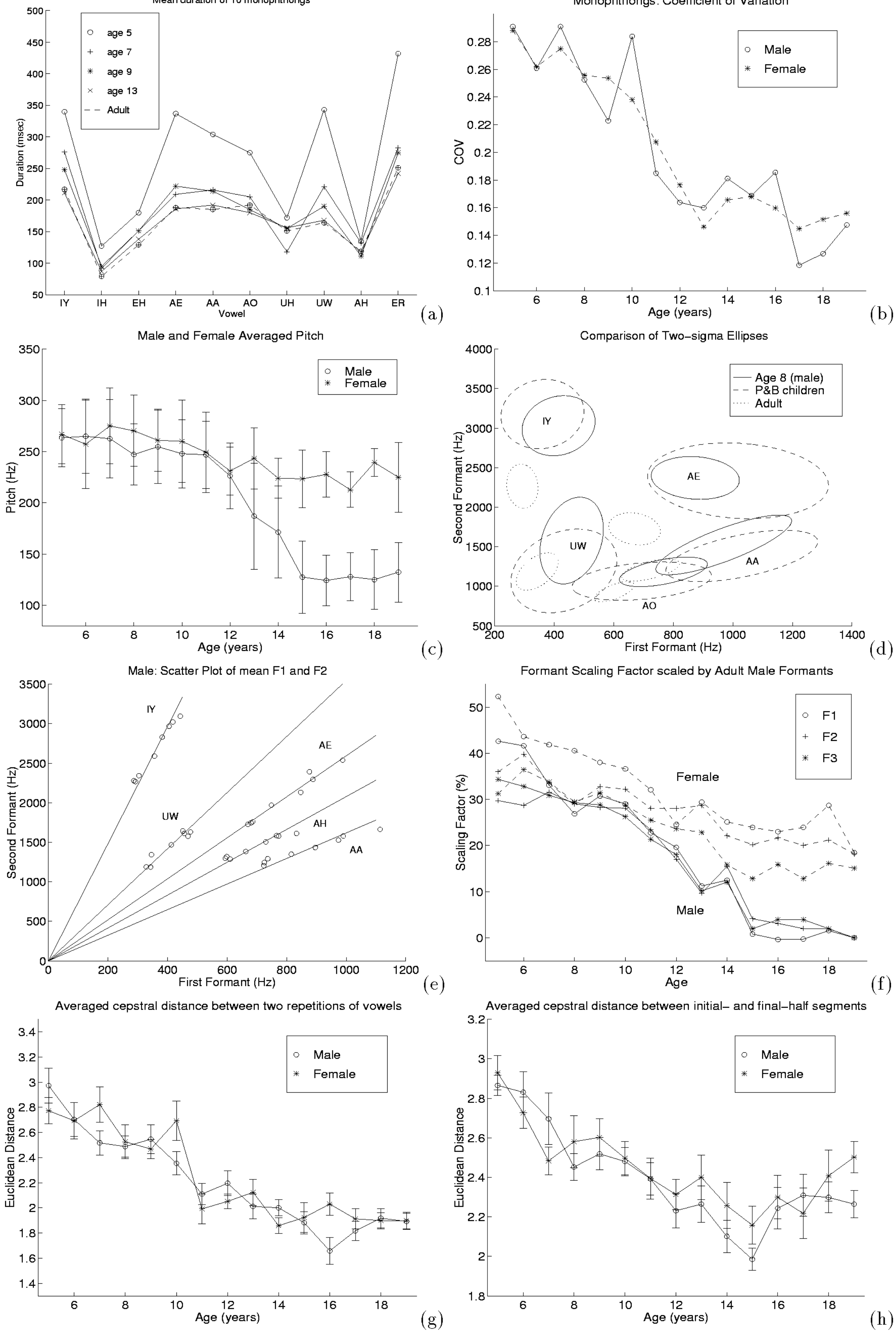

Figure 1: (a) Mean duration of 10 monophthongs. (b) Intra–subject coefficient of variation (age 19 corresponds to adult speakers). (c) Mean pitch of male and female speakers. (d) Two–sigma ellipses for vowels (dotted circles correspond to adult speakers in this database). (e) Plot of mean F1 and F2 of vowels /iy/, /ae/, /aa/, /ao/, and /uw/. (f) Mean formant scaling factors as a function of age and gender. (g) Mean cepstral distance between two repetitions of the same vowel. (h) Mean cepstral distance between the first– and second–half segments of all vowels.