Acoustic Analysis of Preschool Children's Speech

Serdar Yildirim, Shrikanth Narayanan, Dani Byrd and Sonia Khurana
University of Southern California-Integrated Media Systems Center
Speech Analysis and Interpretation Laboratory,
Los Angeles, CA USA
E-mail: [yildirim, shri]@sipi.usc.edu, [dbyrd, skhurana]@usc.edu

ABSTRACT

In this paper, changes in acoustic characteristic of young children speech are investigated as a function of age. We examined fundamental frequency, formant frequencies and vowel durations of ten monophtongal vowels from American-English speaking children, age groups 3 to 6. In this study, we also examined the vowel pronunciation variability of young-aged children. The results show that preschool children speech exhibits high fundamental and formant frequencies and greater variations between subjects for all age groups. Individual vowel durations averaged across all subjects in preschooler follows similar pattern for all age groups. Our statistical analysis shows that age effect is significant on pronunciation variability. Finally, an information-theoretic analysis of developmental changes in the speech signal is presented. The findings based on mutual information between the vowel category and the spectral features correspond well with results of automatic vowel recognition.

1. INTRODUCTION

Acoustic analysis of children speech, especially for younger ages, is challenging in several aspects. Children's speech organs are still developing and their speech-motor control has not yet been fully established. A previous study by Lee et al. [1] showed that children speech exhibits higher pitch and formant frequencies, and longer segmental durations. That study investigated temporal and spectral parameters of children's speech as a function of age and gender. It also showed that spectral and temporal variability in children speech is greater than that in adult's speech.

Age-dependent changes in acoustic parameters, and the greater variability in their values, have serious implications for the design of robust automatic speech recognition (ASR) systems. In fact, most current ASR systems are neither designed nor successful with children especially with preschool age ones. The main underlying reason is the greater acoustic mismatch expected between children speech and the models used for recognition. To understand how to overcome this problem in a principled way, acoustic characteristics of children speech needs to be adequately analyzed and described.

Observations regarding age-dependent scaling of formant frequencies has led to techniques for vocal tract normalization based on frequency warping to help improve recognition performance for children speech [2]. Experiments have shown that even if one were to minimize model level mismatches significantly, there is a gap in the ASR performance levels between those on children speech and adult speech. One of the reasons is that the acoustic spectral feature vectors from children speech contain *less* information about its phonetic class than that of adult's speech [3]. That study also showed that mutual information increases as the signal bandwidth increases for both children and adults. However, amount of information about cepstral features in children speech is less than that for adults for any given bandwidth.

In addition to acoustic factors, another reason for the ASR performance degradation is the child's speaking proficiency [4]. Their work has shown that the error rates for children judged to be poor speakers were four times higher than that for children judged to be good speakers. In this present study, we have included an analysis of the vowel pronunciation variability of preschool children for age groups 3 to 6.

Most previous efforts have targeted ages 5 and older. In this paper we analyze the acoustic parameters of preschool-aged (3-6 years) speech data that we are currently collecting. The analysis also focuses on possible implications to ASR.

The rest of the paper is organized as follows. In Section 2, the speech data corpus analyzed in this work is described briefly. In Section 3, methods and results are given. In Section 4, an information-theoretic analysis of the age-dependent changes in speech spectral parameters is presented. Finally, Section 5 provides a discussion and some conclusions.

2. SPEECH DATA CORPUS

The speech data, analyzed in this work, is a part of the children database that we are currently collecting for a project on children-machine interaction [http://sail.usc.edu/chimp]. The data corpus analyzed contains data from 12 children ages between 3 and 6, 7 females and 5 males. The children were instructed by the computer agent to produce target utterances. Target words are bead (/IY/), bit (/IH/), bet (/EH/), bat (/AE/), but (/AH/), pot (/AA/), ball (/AO/), put (/UH/), boot (/UW/), and bird

ISBN 1-876346-nn-n

(/ER/).

Recordings were made using a high-fidelity microphone connected to a DAT recorder with 44.1 kHz sampling rate and 16-bit resolution. Each waveform file was manually examined and any clipped/truncated data were excluded from the database.

3. SPEECH ANALYSIS

In order to analyze vowel part of speech data, each utterance was phonetically segmented. The beginning and end of each vowel segment and pause period were time marked and saved in a separate label file for each utterance.

3.1 Method

The fundamental frequency and formant frequencies of the ten-monophtongal vowels were estimated using the Praat speech processing software [8]. First, each utterance was down-sampled to 16kHz and formant tracks for the three formants, F1, F2 and F3, were automatically computed using Burg-LPC algorithm with a 10^{th} order linear prediction model. In order to estimate F0 and formant frequencies, the raw data output was smoothed using a 3-point median filter and global median values were computed as a representative F0 and formant frequencies for the track. The segmental duration of vowels were calculated from the corresponding label files that contained the beginning and end time of the vowels.

3.2 Fundamental Frequency

Mean and variation of fundamental frequency, averaged across all vowels and subjects, is shown in Figure 1 as a function of age. Mean F0 values for all ages confirm the trend given in [1]. However, mean pitch value and between subject variations for age 4 was greater than that of other age groups (this may be due to the small sample size of this preliminary study). Also, in our analysis mean values are slightly low for ages 5 and 6 compared to that given in [1]. This difference may be due to dialect difference of subjects between two studies. Still, mean and variation of F0 values are relatively high, as expected, compared to that calculated from adults. A simple factorial analysis of variance (ANOVA) indicates that the effect of age is significant [F=3.946, p=0.01]. Also, multiple a priori comparisons, Bonferroni test with significance level 0.05, show that the group of age 4 exhibits higher averaged F0 than older age groups 5 and 6.

Mean F0 value for each vowel for different ages is shown in Figure 2. It can be easily seen form the figure 2 that age groups 3, 5, and 6 exhibit similar F0 pattern as adults. (Adults data are given from [1].)

We also analyzed effect of vowels on fundamental frequency (averaged across all ages). Analysis of variance indicates that the effect of vowels on F0 is not significant. Also, Multiple comparisons show that no significant difference exists among the vowels.

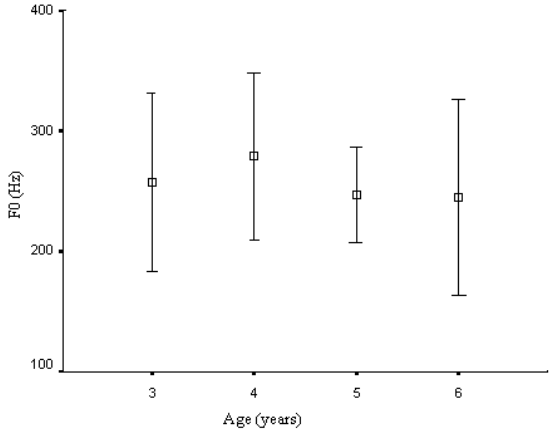

Figure 1. Averaged fundamental frequency. Vertical bars denote between subject variations.

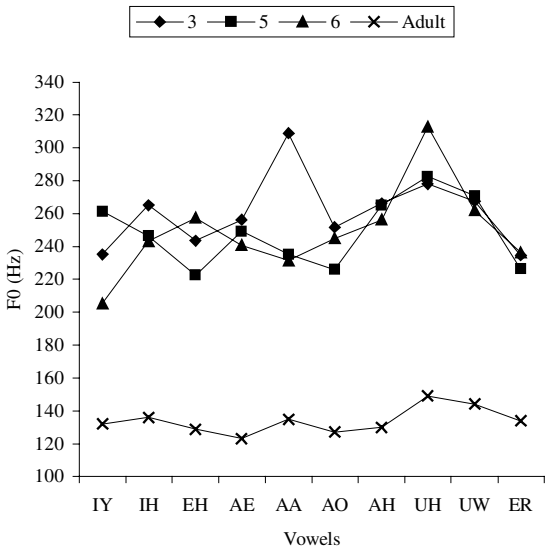

Figure 2. Mean fundamental frequency of individual vowels for preschool children as a function of age.

3.3 Formant Frequencies

Two-sigma ellipses for five vowels are shown in Figure 3 in the F1-F2 space for children speakers' ages 5 to 6 and compared with the data from Lee et al [1]. It can be seen from the figure that within vowel variations are consistent between two studies. The most noticeable difference between two studies is the orientation of vowel /AA/. Even though, there is a dialect difference between two studies, vowel positions are also consistent.

The average first, second and third formant frequencies are given in Table 1 for ages 3, 4, 5, and 6. It can be observed

from the table that young-aged children exhibit high formant frequencies. It is significant that third formant frequencies for age 3 are around or greater that 4kHz.

First three formant frequencies of children and adults have been compared in [5] for various vowels. Their study has shown that third formant frequency on average for children's speech is greater than 4kHz. However, in our study, only third formant frequency values from age group 3 are greater than 4kHz. It possible because of differences in dialects and the vowel set chosen for the analysis.

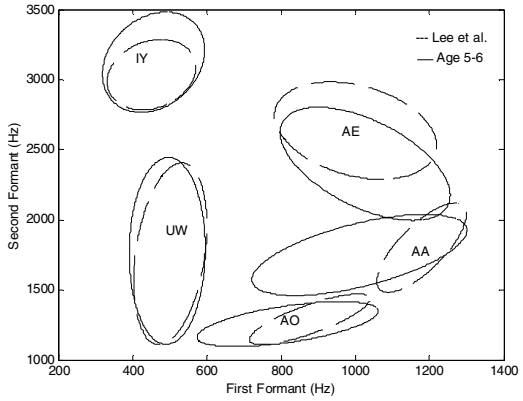

Figure 3. Current formant data are compared with those in Lee et al. [1].

Table 1. Mean of formant frequency values

	F1 (Hz)				F2 (Hz)				F3 (Hz)			
Vowel	3	4	5	6	3	4	5	6	3	4	5	6
IY	541	520	485	437	3335	3163	3226	3054	4054	3671	3804	3680
IH	597	711	481	550	2941	2931	2655	2688	4173	3791	3802	3840
EH	735	803	730	872	2836	2521	2453	2336	4132	3689	3704	3747
AE	917	1157	1194	971	2560	2533	2295	2432	3720	3437	3486	3555
AA	1249	812	1177	956	1876	1828	1864	1710	3790	3499	3768	3601
AO	895	846	771	849	1520	1316	1287	1234	3336	3918	3611	3461
AH	888	838	898	846	1950	1682	2060	1864	3981	3911	3471	3444
UH	676	539	563	663	1659	1353	1730	1633	3939	3581	3489	3423
UW	521	470	528	486	1804	1308	1884	1472	3853	2961	3165	3050
ER	589	579	649	502	1768	1839	2038	1677	2852	2809	2540	2316

3.4 Vowel Duration

Durations of ten vowels were measured using the label files that contain beginning and end time marks of each vowel. Durations averaged across all vowels and subjects in each age group were calculated and analyzed using the SPSS statistical software package. A simple factorial analysis of variance (ANOVA) indicates that the effect of age is not significant within the preschool age groups. Individual vowel durations averaged across subjects for preschool children is given in Figure 4. It is observed from the figure that for all age groups, individual vowel durations averaged across subjects exhibit similar patterns. Despite the low duration values compared to values given [1] for age 5, similar vowel duration patterns are observed.

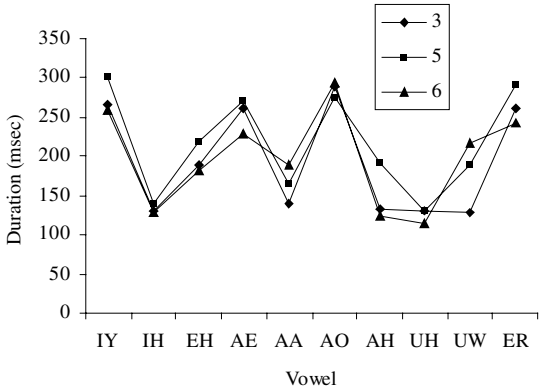

Figure 4. Mean duration of individual vowel averaged across all subjects.

3.5 Vowel Pronunciation Variability

It is reported in [4] that child's speaking proficiency effects ASR performance dramatically. Their work has shown that the error rates for children judged to be poor speakers were four times higher than that for children judged to be good speakers. In this work, we analyzed the vowel pronunciation variability of age groups 3 to 6. A subjective judgment of the vowel quality with respect to the expected target vowel was made. All the speech data from those age groups were rated by a native English speaker as "good", "medium", or "poor". The scores were organized by vowel, subject id, and age groups and analyzed using SPSS statistical package. The results showed that the effect of age is significant [$F=4.125$, $p<0.01$]. Also, multiple a priori compressions with significance level 0.05 indicate that the group of age 3 exhibits significantly less score than age group 6.

4. INFORMATION ANALYSIS

The effects of age and signal bandwidth on speech signal features were analyzed especially motivated by implications to automatic recognition of children's speech using a simple information analysis as in [3]. In [3], the Children Microphone Speech Database [6] was used for the analysis. In Figure 5, mutual information between cepstral feature vector and vowel class are given as a function of bandwidth for different age groups. It is observed form the figure that as bandwidth increases also mutual information increases for all age groups. It is important to note that

ISBN 1-876346-nn-n

information contained in children speech cepstral features never reaches to adult level. The effect of bandwidth change on ASR performance is shown in Figure 6. The vowel classification results correspond well with the mutual information results.

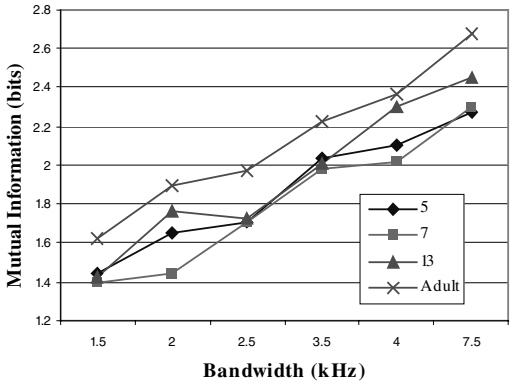

Figure 5. Information changes for different ages (years) with respect to different bandwidths (kHz) for female speakers.

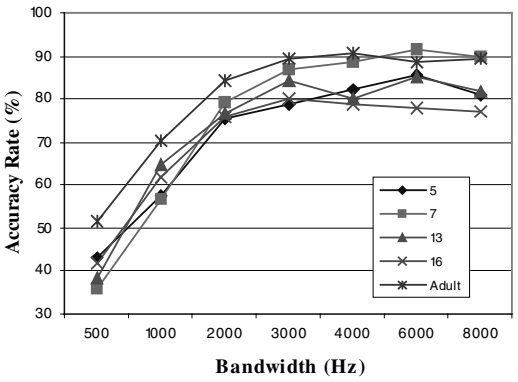

Figure 6. Vowel recognition accuracy results for female speakers.

5. DISCUSSION

In this paper, we analyzed the acoustic characteristics of young children speech as a function of age. We examined fundamental frequency, formant frequencies and vowel durations of ten monophthongal vowels for age groups 3 to 6. Despite the slightly low fundamental frequency averaged across all vowels, the acoustic data and age-dependent trends observed in this study are consistent with the previous studies [1,7]. Despite the dialectal differences between this study and the study by Lee et al. [1], the vowel positions and within vowel variances in the F1-F2 space are consistent. Even though the reduction in magnitude and variability of segmental duration has been observed between age groups 3and 6, significant reduction has not been observed between preschool-aged children. Since previous studies have shown no specific gender effect in younger age groups, gender effects was not considered in this study. Greater variability in acoustic parameters of young children speech may be seen a major obstacle for developing automatic speech recognition and synthesis systems. The effect of this variability was illustrated through an information theoretic analysis.

REFERENCES

[1] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children speech: Developmental changes of temporal and spectral parameters," in *J. Acoust. Soc. Am.*, vol. 105, pp. 1455–1468, 1999.

[2] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children,"' *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 2, pp. 65-78, 2002.

[3] S. Yildirim and S. Narayanan, "An information-theoretic analysis of developmental changes in speech," in Proc. of ICASSP, (Hong Kong), 2003.

[4] Q. Li and M. Russell, "An analysis of the causes of increased error rates in children's speech recognition," in Proc. of ICSLP, (Denver, CO), 2002.

[5] Q. Li and M. Russell, "Why is Automatic Recognition of Children's Speech Difficult?," Eurospeech 2001, Scandinavia.

[6] J.D. Miller, S. Lee, R.M. Uchanski, A. H. Heidbreder, B. B. Richman, and J. Tadlock, "Creating of two children's speech databases," in Proc. of ICASSP, (Atlanta, GA), pp.849–852, 1996.

[7] S. Eguchi, and I. J. Hirsh, "Development of speech sounds in children", Acta Oto-Laryngol. Suppl. 257, pp. 1-51, 1969.

[8] P. Boersma, and D. Weenink, "Praat Speech Processing Software," Institute of Phonetics Sciences of the University of Amsterdam. http://www.praat.org