ANALYZING CHILDREN'S SPEECH: AN ACOUSTIC STUDY OF CONSONANTS AND CONSONANT-VOWEL TRANSITION

M. Gerosa[a,b], S. Lee[c], D. Giuliani[b] and S. Narayanan[c]

([a]) International Graduate School, University of Trento, 38050 Pantè di Povo, Trento, Italy
([b]) ITC-Irst, Centro per la Ricerca Scientifica e Tecnologica, 38050 Pantè di Povo, Trento, Italy
([c]) Department of Electrical Engineering, University of Southern California
Los Angeles, CA 90089, USA

gerosa@itc.it, sungbokl@usc.edu, giuliani@itc.it, shri@sipi.usc.edu

ABSTRACT

This paper presents several acoustic analyses on read speech, collected from 5 adults and 35 children aged 5 to 17 years, focusing on consonants and consonant-vowel transition. Characteristics of consonants such as duration, intra-speaker variability and, for stop consonants, voice onset time are analyzed and compared with results achieved on vowels.

Strong and significant correlation with age is observed for both duration and intra-speaker variability. In fact, younger children show higher phone duration and larger spectral and temporal variability than older children and adults. Voice onset time, on the other hand, is less correlated with age.

Analysis of consonant-vowel transition shows that the duration of the transition and the amount of spectral difference between consonant and vowel are clearly age-dependent. Younger children, in fact, show shorter transition duration and larger spectral difference between consonant and vowel in the consonant-vowel pair.

1. INTRODUCTION

It is well known that acoustic and linguistic characteristics of children's speech are widely different from those of adult speech [1, 2, 3]. For example, children's speech is characterized by higher pitch and formants frequencies with respect to adults' speech. Furthermore, characteristics of children's speech vary rapidly as a function of age due to the anatomical and physiological changes occurring during a child's growth and because children become more skilled in co-articulation with age.

Much has been done in the past in analyzing the acoustic differences between children's and adult speech, with a particular focus on vocal tract length and its influence on pitch and formant frequency values [1]. Understanding the developmental changes in children's speech can help devise strategies for dealing with the acoustic mismatch between different age groups, for example in applications such as Automatic Speech Recognition (ASR) [4] and in early literacy and reading assessment [5].

Overall, the majority of the previous efforts in children's speech analysis has dealt with vowel duration, pitch, and formants with little or no work on consonant analysis. In this work, a subset of the CID corpus [6], already used in [1] for a comprehensive analysis on vowels, was used to analyze consonants, considering features such as Voice Onset Time (VOT), duration and spectral and temporal variability, and consonant-vowel (CV) transitions.

CV transition was analyzed with the aim of studying the effect of changes in co-articulation occurring during growth. In fact, one of the most important causes of speech variation is co-articulation, in which the realization of a particular phone is affected by its neighbors. Some past work analyzed developmental changes in co-articulation [2, 3], but all studies were focused only on few sounds, usually fricatives, and lacked a systematic study on several different phonetic classes.

The paper is organized as follows. The speech corpus used in this work is described in Section 2. Section 3 presents the results on the analyses performed on consonant duration, VOT and intra-speaker variability. Section 4 shows the analyses on CV transitions. Finally, discussion about the results and concluding remarks are given in Section 5 that concludes the paper.

This work was supported in part by the National Science Foundation and was carried out while the first author was at University of Southern California, Los Angeles.

2. SPEECH CORPUS

The CID corpus consists of read speech collected from 436 children, aged from 5 to 18, and from 56 adult speakers [6]. Speech was acquired at 20 kHz, with 16 bit accuracy, using a high-fidelity microphone. The signals were downsampled to 16 kHz before being analyzed. Only a subset of this database was analyzed in this work. Data from five speakers, 3 females and 2 males, for ages 5, 7, 9, 11, 13, 15, 17 and from five adult speakers were considered, for a total of 40 subjects.

The speech material analyzed in this paper consisted of repetitions of five phonetically-rich and meaningful sentences. Each sentence was uttered two times by each speaker. 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." and 5 "Did you like the zoo this spring?". Prior to the recording session, any target utterances that the speakers, mostly 5 years olds, had difficulty reading were identified and elicited through imitation of a sample prerecorded by a female speech pathologist.

Manual segmentation at the phone level was performed in the following way. First, age-dependent HMMs were trained on the CID corpus and then phone level segmentation was performed using the transcription of each utterance. Optional insertion of "silence" was allowed at the beginning and the end of each utterance and between words. After automatic segmentation, each segment representing a stop consonant was additionally segmented in two parts. The first part was marked as the stop closure, while the second part as the stop burst. Each utterance was then analyzed by a native speaker of English with good phonetics knowledge. The annotator modified the boundaries of the phonetic segmentation in order to correct segmentation errors.

Without considering the boundary between the closure and burst part of each stop consonant, the mean segmentation difference between the automatically computed and the manually measured values was 15.3 msec with a standard deviation of 27.3 msec. The percentage of automatically generated boundaries located within 20 msec from the manual boundaries was 74%.

3. SPEECH ANALYSIS

3.1. Temporal characteristics

In this work, we compared durations of vowels and two different classes of consonants. The mean phone duration was computed by first averaging phone durations over all phones of a certain phonetic category of each speaker and then across speakers in each age group. Duration statistics were computed by exploiting the phone-level manual segmentation produced as explained in Section 2. Figure 1 reports the mean phone duration of vowels for children of different ages and adults, while Figure 2 reports the same analysis on fricative and stop consonants. For duration of stop consonants, we considered the closure and burst parts together.

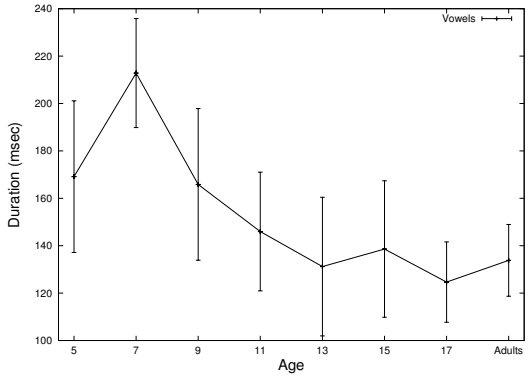

Fig. 1. *Mean duration of vowels (msec) per age. Vertical bars denote inter-speaker variability (standard deviation).*

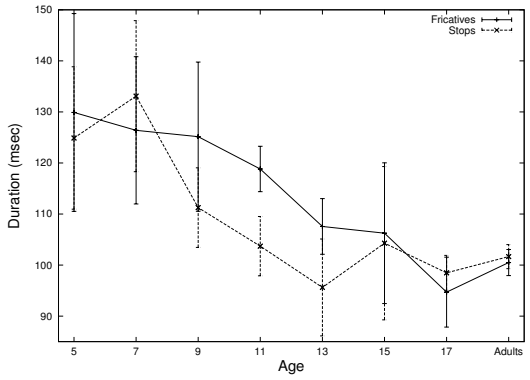

Fig. 2. *Mean duration of fricative and stop consonants (msec) per age. Vertical bars denote inter-speaker variability (standard deviation).*

It can be noted that mean phone duration varies with age and older age groups exhibit shorter mean phone duration. Analysis of variance (ANOVA) showed that effect of age is significant with $p < .001$ in all three cases. However, we have to point out that mean phone durations reported here are likely affected by reading ability. The significant difference between values obtained for vowels uttered by children of age 5 and 7 can be explained by the fact that the latter read a set of sentences while for the former, speech was elicited through imitation of a sample recorded by an adult. Vowels and consonants present the same trend, even if, as expected, the duration of vowels is higher than that of consonants. Moreover, while fricative consonants decrease in duration is almost linear with age, for vowels and stop consonants it is concentrated between ages 7 and 13.

3.2. Intra-speaker variability

Intra-speaker variability is a measure of the maturity of speech motor control. In this work, we characterized intra-speaker variability as the temporal and spectral difference between corresponding phones in two repetitions of the same sentence. As described in Section 2, five phonetically rich sentences were uttered two times by each speaker. We considered the two repetitions of the same sentence uttered by a given speaker and measured the spectral and temporal difference phone by phone, from the beginning to the end of each pair of utterances.

To perform the spectral analysis, the speech signal was first blocked into frames of 20 ms duration (with 50% frame overlapping), then each speech frame was parameterized into 12 mel frequency cepstral coefficients (MFCCs). Each MFCC was scaled with the inverse of the standard deviation computed over all data. The mel cepstrum distance between two speech segments was computed by first computing the average MFCC vector for each segment, and then taking the Euclidean distance between the two vectors.

Figures 3 and 4 show the temporal and spectral difference computed on consonants and vowels, averaged over all phones of a given speaker and then across all speakers in each age group.

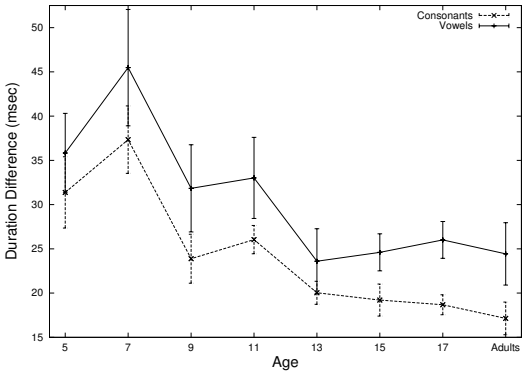

Fig. 3. *Mean duration difference between corresponding phones in two repetitions of the same sentence. Vertical bars denote inter-speaker variability (standard deviation).*

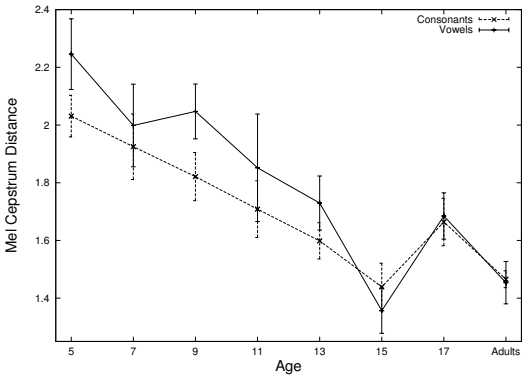

Fig. 4. *Mean mel cepstrum distance between corresponding phones in two repetitions of the same sentence. Vertical bars denote inter-speaker variability (standard deviation).*

Observing Figures 3 and 4, it is clear that intra-speaker variability, both spectral and temporal, decreases with age. Analysis of variance showed that effect of age is significant in all the four curves with $p < .001$. Imitated speech produced by speakers of age 5 shows

shorter temporal variability but not spectral variability. One possible explanation is that while repeating speech uttered by an adult, children are able to imitate his/her temporal pattern, their articulation control remains still uncertain.

Another interesting characteristic is that the minimum for spectral variability is observed for children of age 15. This behavior was already observed for vowels in [1], however the reason is not clear. This phenomenon could be associated with the learning process or it may be that the articulation control capability peaks during teenage years.

3.3. Voice Onset Time

The Voice Onset Region (VOR) of a stop is the region of unvoiced speech that starts after the closure part of the stop (stop release area) and ends just before the onset of the voicing of the vowel. The length of the VOR is called voice onset time, and it reflects the degree of coordination between articulation and voicing. This feature is generally ignored in the common fixed frame length speech analysis, although it is known that VOT helps the listener in recognizing which stop is being produced. We exploited the manual segmentation of the closure and burst part of stop consonants to measure VOT. In particular, we considered VOT for consonant /P/ in the word "pen" and consonant /T/ in word "tall". Each child uttered the word two times, for a total of 10 examples for each age. In Figure 5, the results on the VOT of /P/ and /T/ are shown.

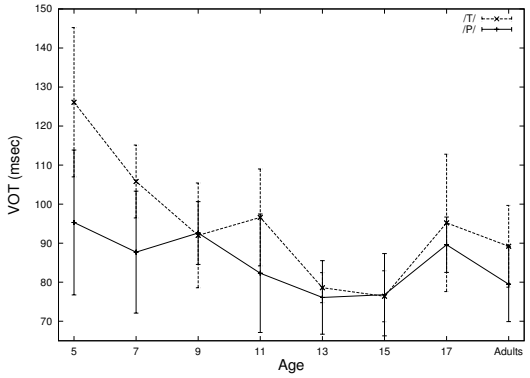

Fig. 5. *Mean VOT for /T/ in "tall" and /P/ in "pen' for each age. Vertical bars denote inter-speaker variability (standard deviation).*

It can be noted that the aspiration after the burst of /T/ and /P/ make the VOT particularly long. VOT decreases up to age 15 for both the considered cases, but ANOVA analysis showed that this decrease is significant ($p < .005$) for /T/ but not significant for /P/ ($p = .18$). VOT seems to be less correlated with age than the general phone duration.

4. CV TRANSITION CHARACTERISTICS

One of the most important causes of acoustic variation in speech is co-articulation, in which the realization of a particular phone is affected by its neighbors. We investigated the effect of co-articulation in the transition between consonant and vowel, in particular the duration of the transition and the mel cepstrum distance between consonant and vowel in a given CV pair.

Here, we assume that if the spectral characteristics of speech changes slowly in the CV pair, then the amount of co-articulation is higher, while if the spectral transition is abrupt the amount of co-articulation is more limited.

In general, the spectral change between two phonemes is a function of: 1) inherent spectral distance between two phones in isolation, 2) local articulatory movement velocity and speaking rate and 3) target attainment (under- or overshoots). For a given speaker and a target syllable, these are related to the agility in articulation and skill of making minimal contrast between two phones without loosing phonemic identities. Therefore, we think the spectral change between two phonemes can be correlated to the maturity of co-articulation skill.

With this assumption, we measured the rate of CV transition in the following way. First, we computed the mean MFCC vectors, $\bar{\boldsymbol{x}}_c$ and $\bar{\boldsymbol{x}}_v$, for consonant and vowel in a given CV occurrence, by averaging feature vectors over all frames in each phone. Then we computed, for each frame i in the CV occurrence, the function:

$$f_{cv}(i) = d(\bar{\boldsymbol{x}}_c, \boldsymbol{x}_i) - d(\bar{\boldsymbol{x}}_v, \boldsymbol{x}_i)$$

where $\boldsymbol{x}_i$ denotes the feature vector of frame i and $d(\cdot,\cdot)$ denotes the Euclidean distance between two feature vectors. In Figure 6, this function is plotted for one particular example of the CV pair corresponding to the word "she", that we denote with /SH/-/IY/ using Darpa symbols to identify the phones.

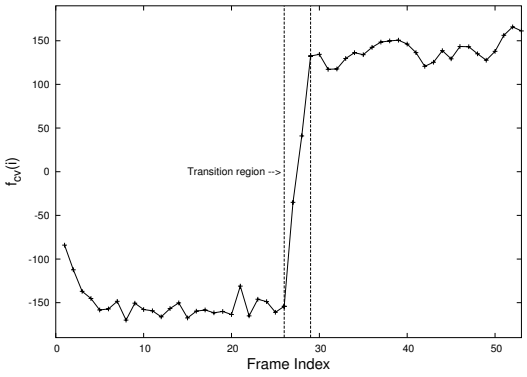

Fig. 6. *$f_{cv}(i)$ values computed on a frame-by-frame basis for the transition /SH/-/IY/ in an instance of the word "she".*

Using this function, we computed the relative duration of the transition, with respect to the length of the CV pair. Transition duration (i.e. length of the transition region depicted in Figure 6), was estimated by using an heuristic algorithm that selected frames showing an increase in f_{cv}, with respect to their preceding frame, greater than a fixed threshold. In CV transitions with stop consonants, only the burst part of the stop was considered. Figure 7 reports the result obtained for /SH/-/IY/, averaging relative durations of the transitions first across the two instances of word "she" uttered by each speaker and then across all the speakers in each age groups.

13 different CV pairs in the CID corpus were analyzed. In Table 1, the correlation coefficient (r) between relative duration and age for each CV pair is reported. CV pairs were grouped depending on the consonant phonetic category, following the classification reported by [7].

It can be noted that for each CV pair the relative duration of the transition shows some correlation with age. ANOVA performed on the groups of CV pairs corresponding to the three phonetic categories shows significance level of $p < .001$, $p < .01$ and $p < .05$ for fricatives, stops and liquids, respectively. Considering all the different CV pairs together the level of significance is $p < .001$ and the correlation coefficient with age is r=0.87.

Another aspect of the CV pair we measured is the spectral difference between consonant and vowel. We measured this feature as the Euclidean distance between the two mean MFCC vectors obtained by averaging MFCCs across all frames in each phone. Then, we averaged the Euclidean distance over all CV pairs of a certain category for each speaker, and across all speaker in each age group. Table 2 reports the correlation coefficients between age and the mel cepstrum distance computed on the CV pairs grouped in the three

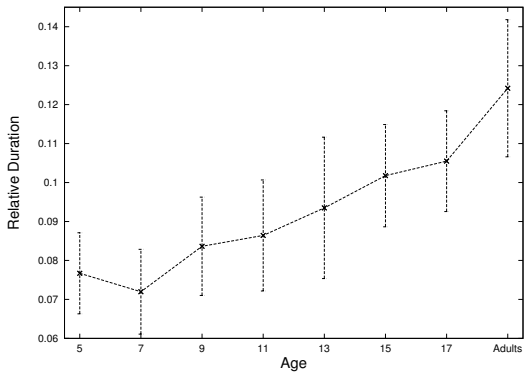

Fig. 7. *Relative duration of the transition between phones /SH/ and /IY/ in word "she". Vertical bars denote inter-speaker variability (standard deviation).*

Category	CV pair	word	r
Fricatives	/CH/-/AH/	**Chu**ck	0.57
	/JH/-/AE/	**ja**m	0.78
	/SH/-/IY/	**she**	0.95
	/S/-/IY/	**see**ms	0.73
	ALL	-	0.96
Stops	/P/-/EH/	**pe**n	0.36
	/T/-/AO/	**ta**ll	0.49
	/T/-/OW/	**toa**st	0.64
	/D/-/IH/	**di**d	0.68
	ALL	-	0.89
Liquids	/HH/-/IY/	**he**	0.54
	/HH/-/AE/	**ha**s	0.58
	/R/-/AO/	st**raw**berry	0.43
	/R/-/IY/	strawber**ry**	0.42
	/L/-/UW/	b**lue**	0.62
	ALL	-	0.62

Table 1. Correlation with age (r) of the relative duration of the transition for several CV pairs.

phonetic categories and on all the CV pairs together, as well as the significance levels obtained performing ANOVA.

CV pair category	r	p
Fricatives	0.88	.001
Stops	0.50	-
Liquids	0.64	.05
ALL	0.80	.001

Table 2. Correlation with age (r) and significance level (p) of mel cepstrum distance between consonant and vowel in a given CV pair.

Correlation between mel cepstrum distance and age is not as strong as the one obtained for the duration of the CV transition, but still present. In fact, ANOVA measures show that the effect of age is significant for fricative ($p < .001$), somewhat significant ($p < .05$) for liquids and not significant for stop consonants. However, when considering all CV pairs together, the correlation with age is strong, r=0.8, and significant ($p < .001$).

5. DISCUSSION AND CONCLUSIONS

In this paper, results of several acoustic analyses on children's and adult read speech focusing on consonants and CV transition have been presented. Analyses of duration and intra-speaker variability performed on consonants and vowels showed that these features, when studied in relation with age, have a similar trend for the two phonetic classes.

In duration analysis, the relative reduction of phone duration for consonants and vowels, from age 7 to age 17, is about 25% for consonants and 41% for vowels. This means that the ratio between vowel and consonant duration is higher for younger children, about 1.64 for children of age 7 and 1.40 for age 9, than for older children and adults, about 1.3 for subject of age 11 and older.

Analysis on intra-speaker variability showed that spectral and temporal variability decrease with age, and vowel variability is higher than consonant variability. However, even if the absolute temporal variability for consonants is lower, if we compute the relative variability with respect to the mean phone duration, reported in Figures 1 and 2, we note that there is no significant difference between the values for consonants and vowels.

VOT analysis confirmed the decrease in duration with age, even if it is not as marked as the one observed in Section 3.1. This is an evidence that the coordination between consonantal articulation becomes more effective as children grow older and skill in the speech motor control improves.

CV transition analysis showed an interesting increase with age of the relative duration of transitions. As explained in Section 4, this may imply that the effect of co-articulation is less marked for younger children than for adults.

The investigation on spectral difference between consonant and vowel in each CV group pointed out that the mel cepstrum distance between consonant and vowel in a given CV pair increases with age. This is consistent with past results that show that the confusability between different phones is higher for children than for adults, as spectral characteristics of different phones uttered by adults are more distant from one another than children's.

6. REFERENCES

[1] S. Lee, A. Potamianos, and S. Narayanan, "Acoustic of children's speech: Developmental changes of temporal and spectral parameters," *Journal of Acoust. Soc. Amer.*, vol. 105, no. 3, pp. 1455–1468, March 1999.

[2] S. Nittrouer and D.H. Whalen, "The Perceptual Effects of Child-Adult Differences in Fricative-Vowel Coarticulation," *Journal of Acoust. Soc. Amer.*, vol. 86, no. 4, pp. 1266–1276, Oct. 1989.

[3] R.S. McGowan and S. Nittrouer, "Differences in Fricative Production Between Children and Adults: Evidence from an Acoustic Analysis of /f/ and /s/," *Journal of Acoust. Soc. Amer.*, vol. 83, no. 1, pp. 229–236, Jan. 1988.

[4] M. Gerosa, D. Giuliani, and F. Brugnara, "Speaker Adaptive Acoustic Modeling with Mixture of Adult and Children's Speech," in *EUROSPEECH*, Lisbon, Portugal, Sept. 2005, pp. 2193–2196.

[5] A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P. Price, E. Anderson, S. Narayanan, and A. Alwan, "Tball data collection: the making of a young children's speech corpus," in *EUROSPEECH*, Lisbon, Portugal, Sept. 2005, pp. 1581–1584.

[6] J.D. Miller, S. Lee, R.M. Uchanski, A.H. Heidbreder, B.B. Richman, and J. Tadlock, "Creation of Two Children's Speech Databases," in *Proc. of ICASSP*, Atlanta,GA, May 1996, pp. 849–852.

[7] J. J. Odell, "The Use of Context in Large Vocabularry Speech Recognition," *Ph. D. Thesis, Cambridge University, Cambridge*, 1995.