A Review of the Acoustic and Linguistic Properties of Children's Speech

Alexandros Potamianos
Dept. of Electronics and Computer Engineering
Technical University of Crete
Chania 73100, Greece
Email: potam@telecom.tuc.gr

Shrikanth Narayanan
Speech Analysis and Interpretation Laboratory (SAIL)
Viterbi School of Engineering
University of Southern California
Los Angeles, CA 90089, USA
Email: shri@sipi.usc.edu

***Abstract*— In this paper, we review the acoustic and linguistic properties of children's speech for both read and spontaneous speech. First, the effect of developmental changes on the absolute values and variability of acoustic correlates is presented for read speech for children ages 6 and up. Then, verbal child-machine spontaneous interaction is reviewed and results from recent studies are presented. Age trends of acoustic, linguistic and interaction parameters are discussed, such as sentence duration, filled pauses, politeness and frustration markers, and modality usage. Some differences between child-machine and human-human interaction are pointed out. The implications for acoustic modeling, linguistic modeling and spoken dialogue systems design for children are discussed.**

I. Introduction

Developmental changes in speech production introduce age-dependent spectral and temporal variabilities in the speech produced by children. Such variabilities pose challenges for spoken dialogue system design for children. Early spoken dialogue application prototypes that were specifically aimed at children included word games for pre-schoolers [27], aids for reading [22] and pronunciation tutoring [26]. Recently a number of systems have been implemented with advanced spoken dialogue interfaces, multimodal interaction capabilities and/or embodied conversational characters [23], [13], [4], [5]. Data collected from these systems as well as new available corpora [2], [28], [3] have improved our understanding of verbal child-machine interaction.

In this paper, we review some of the acoustic and linguistic characteristics of read and spontaneous speech of children ages 6 years and up. The main acoustic analysis results are from [18] but also corroborating evidence from the literature is presented. Then the acoustic and linguistic properties of spontaneous child-machine interaction is reviewed. Finally we conclude with the applicability of these results to speech recognition and spoken dialogue system design.

II. Children corpora

Most of the databases of children recordings focus on the 6-18 age group (or a subset thereof) where collection conditions can be more easily controlled and the subjects are collaborating. Examples of corpora mostly used for acoustic analysis and modeling are the American English CID children corpus [18], the KIDS corpus [9], the CU Kids' Audio Speech Corpus [13] and the PF-STAR corpus available in the following languages: British English, Italian, German and Swedish [2].

As far as spontaneous speech is concerned, including child-machine spoken dialogue interaction or multimodal interaction a handful of corpora has been recently collected and analyzed. In [4], the NICE fairy-tale corpus is presented, where children use open-ended spoken dialogue to interact with animated characters in a game setting. In [3], a child-robot interaction corpus is presented; children interacted with an AIBO robot is open-ended scenarios. In [23], a corpus collected in a Wizard-of-Oz scenario, were children used speech to play a computer game and interact with animated characters on screen is presented and analyzed. In [28], a corpus of child-machine interaction via a multimodal voice and pen interface was collected and analyzed.

More corpora will be made available as the interest in multimodal spoken dialogue systems for children users increases. Another trend in data collection for children is to collect and quantitatively analyze the acoustic and linguistic characteristics of very young children (ages 2-6).

III. Analysis of children's speech

The spectral and temporal characteristics of children's speech are highly influenced by growth and other developmental changes and are hence different from those of adult speakers. These differences are attributed mainly to anatomical and morphological differences in the vocal-tract geometry, less precise control of the articulators and a less refined ability to control suprasegmental aspects such as prosody.

In a key study by Eguchi and Hirsh [8], and later summarized by Kent [16], age-dependent changes in formant fundamental frequency measurements of children speakers ages three to thirteen were reported. Important differences in the spectral characteristics of children voices when compared to those of adults include higher fundamental and formant frequencies, and greater spectral *variability* [8], [16], [18]. Parametric models for transforming vowel formant frequency of children speakers to the adult speaker space (vowel formant frequency normalization) were considered in [12], [21], [25]. Similarly, a detailed comparison of temporal features and speech segment durations for children and adult speakers can be found in [17], [18]. Again, distinct age-related differences

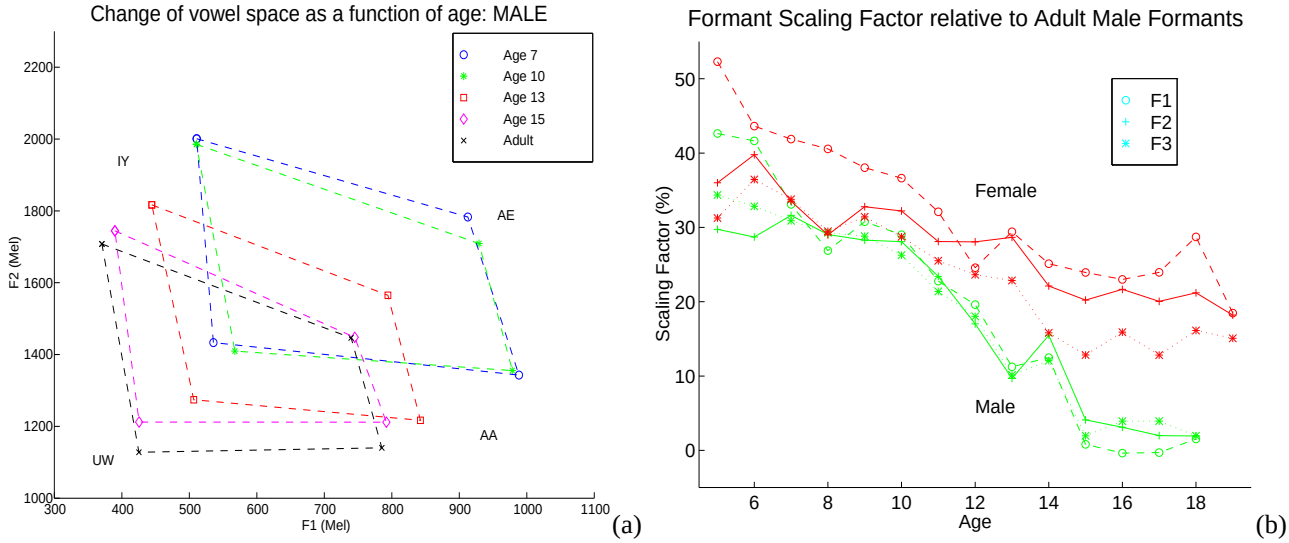

Fig. 1. (a) Changes in F1-F2 vowel space as a function of age. The vowel space boundaries are marked by average formant frequency values for the four point vowels /AA, IY, UW, AE/ for the age groups: 7, 10, 13, 15 and adults. (b) Scaling factor variation in first three formant frequencies with respect to age for male and female children. Scaling was with respect to average values for adult males (from [25]).

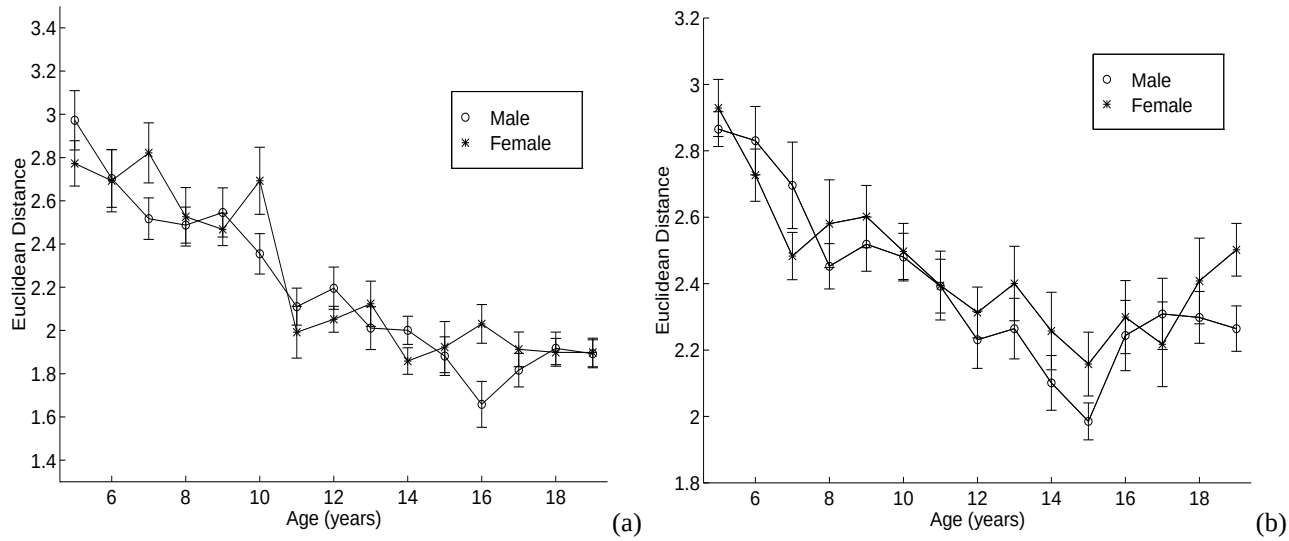

Fig. 2. Intra-speaker variability as a function of age: (a) Mean cepstral distance between the two repetitions of the same vowels and (b) Mean cepstral distance between the first– and second–half segments within the same vowel realization (from [25]).

were found. On average, the speaking rate of children is slower than that of adults. Further, children speakers display higher variability in speaking rate, vocal effort, and degree of spontaneity.

Many of the early acoustic studies were somewhat limited in terms of the size of the database especially the number of subjects. In a related study, variations in the temporal and spectral parameters of children's speech were investigated using a comprehensive speech data corpus (23454 utterances) obtained from 436 children ages between 5 and 18 years and 56 adults [18]. Key findings from that study that focuses on the acoustic properties of the vowels, including results on formant scaling are summarized in the next section. For recent work on acoustic properties of consonants see [11].

A. Age-dependency in acoustic characteristics

To obtain insights into age-dependent behavior in the magnitude and variance of the acoustic parameters, measurements of spectral and temporal parameters were made through a detailed analysis of the American English vowels [18]. Recent work on the analysis of the acoustic characteristic of children speech in other languages provided similar results, e.g., see [10] for Italian. Results showed a systematic decrease in the values of the mean and variance of the acoustic correlates such as formants, pitch and duration with age, with their values reaching adult ranges around 13 or 14 years. A specific result that is especially relevant for speech modeling is the scaling behavior of formant frequencies with respect to age. As can be seen from Fig. 1(a), the vowel space (boundaries marked

by the four-point vowels /AA, IY, UW, AW/ in the F2-F1 plane plotted in mel frequency scale) changes with increasing age in an almost linear fashion. The movement of the vowel quadrilateral is in the direction toward smaller F2-F1 values with increasing age corresponding to the lengthening of the vocal tract associated withe growth. Also, it can be noticed that the the vowel space becomes more compact with increasing age indicating a decreasing trend in the dynamic range of the formant values. The changes in the F2-F1 values are almost linear. A more detailed account of the scaling behavior can be obtained by plotting the variation in the formant scaling factors (calculated as a ratio of average formant frequency values for a specific age group to the corresponding values for adult males). The plots in Fig. 1(b) show a distinct and an almost linear scaling of each of the first three formant frequencies with age. The scaling trend for females and males is similar until puberty suggesting underlying differences in anatomical growth patterns. Moreover, the first three formants scale more uniformly for males. Formant frequencies of females, on the other hand, show a more nonlinear scaling trend for the various formants especially after puberty.

The intra-speaker variability (i.e., within subjects) was larger for young children, especially for those under 10 years. Fig. 2 shows a decreasing trend in intra-subject variability with age in terms of cepstral distance measures of variability both within a token and across two repetitions. It is generally believed that both the acoustics and linguistic correlates of children speech are more variable than those of adults. For example, the area of the F1-F2 formant ellipses is larger for children than for adults for most vowel phonemes [8] and children speech contains more disfluencies and extraneous speech [27]. An important point is that such results are highly dependent on whether the data was read or spontaneous speech.

B. Spontaneous Speech and Spoken Dialogue Interaction

Some insights regarding the acoustic and linguistic characteristics of children's spontaneous speech can be obtained from the results in [24]. The analysis is based on data from a Wizard of Oz study using 160 children playing a voice-activated computer computer game. The average sentence duration was about 10% longer for younger children. As a result, the speaking rate for the 11-14 year-olds was about 10% higher than for the younger group which is in agreement with the results on read speech [18]. An important aspect of spontaneous speech the prevalence of disfluencies. Disfluencies and hesitations in the speech data were analyzed as a function of age and gender. Mispronunciations, false-starts, (excessive) breath noise and filled pauses (e.g., um, uh) were manually labeled for a subset of the data (22422 utterances). About 2% of the labeled utterances contained false-starts and 2% contained (obvious) mispronounciations. Breathing and filled pauses were found in 4% and 8% of the utterances, respectively. While no gender dependency was found for any of the disfluency measures, there was a distinct age dependency. The frequency of mispronunciations was almost twice as high for the younger (8-10 years) age group than for the older group (11-14 years). Breathing noises occurred 60% more often for younger children. Surprisingly, this trend was reversed for filled pauses which occurred almost twice as often for the 11-14 age group. In [1], politeness and frustration markers were analyzed on this database. Younger children used politeness markers more commonly and expressed frustration verbally more often than older children.

In [4], significant differences in the duration and language usage where found in child-machine dialogue compared to human-human dialogue. Specifically children ages 8-15 communicated with fairy-tale characters in a computer game scenario, using shorter utterances, slower speaking rate and much less filled pauses, filler words and phrases, compared to human-human dialogue. In [28], the multimodal integration patterns of children ages 7-10 were investigated for a speech and pen interface. It was found that the modality usage was similar between children and adults, although children tend to use both input modes simultaneously rather than sequentially.

IV. Implications to Spoken Dialogue Systems

There are several implications that the acoustic characteristics mentioned above have for automatic speech recognition (ASR) for children. The main goal of the ASR feature extraction stage is to decompose the speaker-dependent information (e.g., pitch) from the phoneme-dependent information (e.g., formants) and retain the latter. This task is more difficult for children voices because the fundamental frequency and the formant bandwidths are of comparable magnitude. As a result, speaker dependent information exists in the feature vectors derived from children speech which, in turn, results in degradation of the performance of the classification process.

Another major challenge in acoustic modeling for ASR is the spectral and temporal variability in children's speech. Increased variability in formant values results in greater overlap among phonemic classes for children than for adult speakers, and makes the classification problem inherently more difficult. Further, the range of values for most acoustic parameters is much larger for children than for adults. For example, five-year old children have formant values up to 50% higher than male adults [19]. For example, the difficulty for spectral-feature based pattern classification due to increased dynamic range of acoustic parameters is illustrated in the F1-F2 formant space (see Fig. 1(a)). The size of the phonemic classes (represented by the area of the ellipses in the F1-F2 plot) for children speakers ages 5-16 is much larger than for adults, which results in significant overlap among classes (in the F1-F2 space). The combination of a large acoustic parameter range and increased acoustic variability can seriously degrade ASR performance. In [6], [25], speaker normalization procedures and age-dependent acoustic modeling are used to reduce variability and increase resolution between classes. It can be seen from these results, that very good speech recognition performance levels can be achieved for children older than 10 years of age, and good performance is obtained for ages 6-9.

Finally, other issues in ASR for children relate to the effects of spontaneity and greater linguistic variability of children's speech (that creates large amounts of extraneous speech) and the associated ASR interface issues. Although disfluencies and hesitation phenomena occur more frequently in children than in adults, our experiments showed that ASR performance does not suffer significantly due to these effects, hence requiring no special acoustic modeling strategies. However as children are faced with harder tasks, e.g., e-learning, these effects might become more prominent [13]. Finally, note that children differ from adults not only at the acoustic and linguistic level but also their requirements as users at the application and interface level are different, e.g., adults exploit more and explore less than children[7]. Interface and application design issues for children users at equally important but are beyond the scope of this review.

V. Discussion

We know that children speech is quite different from adult speech both in terms of absolute values and variability of acoustic and linguistic correlates. However, despite these differences that make acoustic and linguistic modeling for children more challenging than for adults, efficient algorithms now exist for modeling children speech that provide good performance. Further research is necessary to improve these algorithms and apply them to spoken dialogue system design for children.

We have only started to formally investigate the acoustic, linguistic and interaction patterns of children when interacting with computers, toys or animated characters. Further research is needed to better understand spoken and multimodal child-machine interaction, as well and formally analyze children speech in very young ages (2-5 years of age).

References

[1] S. Arunachalam, D. Gould, E. Andersen, D. Byrd, and S. Narayanan, "Politeness and frustration language in child-machine interactions," in *Proc. European Conf. on Speech Communications and Technology*, 2001.

[2] A. Batliner et al, "The PF-STAR Children's Speech Corpus," in *Proc. of Interspeech*, (Lisbon, Portugal), 2005.

[3] A. Batliner, C. Hacker, S. Steidl, E. Noth, S. D'Arcy, M. Russel and M. Wong, "You stupid tin box" - children interacting with the AIBO robot: a cross-linguistic emotional speech corpus," in *Proc. of the 4th Intern. Conf. of Language Resources and Evaluation*, (Lisbon, Portugal), 2004.

[4] L. Bell, J. Boye, J. Gustafson, M. Heldner, A. Lindstrom, and M. Wiren, "The Swedish NICE Corpus Spoken dialogues between children and embodied characters in a computer game scenario," in *Proc. of Interspeech*, (Lisbon, Portugal), 2005.

[5] L. Bell and J. Gustafson, "Children's convergence in referring expressions to graphical objects in a speech-enabled computer game," in *Proc. of Interspeech*, (Antwerp, Belgium), 2007.

[6] D. C. Burnett and M. Fanty, "Rapid unsupervised adaptation to children's speech on a connected-digit task," in *Internat. Conf. Speech Language Processing*, (Philadelphia, PA), Oct. 1996.

[7] J. Cassell and K. Ryokai, "Making Space for Voice: Technologies to Support Children's Fantasy and Storytelling," *Personal Technologies*, vol. 5, 2001.

[8] S. Eguchi and I. J. Hirsh, "Development of speech sounds in children," *Acta Oto-Laryngologica*, vol. Supplementum 257, pp. 1–51, 1969.

[9] M. Eskernazi, "KIDS: A database of children's speech," *Journal of the Acoustical Society of America*, vol. 100, 1996.

[10] M. Gerosa, D. Giuliani and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, to appear, 2007.

[11] M. Gerosa, S. Lee, D. Giuliani and S. Narayanan, "Analayzing Children's Speech: an Acoustic Study of Consonants and Consonant-Vowel Transition," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*, (Toulouse, France), May 2006.

[12] U. G. Goldstein, *An Articulatory Model for the Vocal Tracts of Growing Children*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, June 1980.

[13] A. Hagen, B. Pellom, and R. Cole, "Children's speech recognition with application to interactive book and tutors," in *Proc. ASRU Workshop*, 2003.

[14] W. F. Katz, C. M. Beach, K. Jenouri, and S. Verma, "Duration and fundamental frequency correlates of phrase boundaries in productions by children and adults," *Journal of the Acoustical Society of America*, vol. 99, pp. 3179–3191, 1996.

[15] W. F. Katz, C. Kripke, and P. Tallal, "Anticipatory coarticulation in the speech of adults and young children: Acoustic, perceptual and video data," *Journal of Speech and Hearing Research*, vol. 34, pp. 1222–1232, Dec. 1991.

[16] R. D. Kent, "Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies," *Journal of Speech and Hearing Research*, vol. 19, pp. 421–447, 1976.

[17] R. D. Kent and L. L. Forner, "Speech segment durations in sentence recitations by children and adults," *Journal of Phonetics*, vol. 8, pp. 157–168, 1980.

[18] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *Journal of the Acoustical Society of America*, pp. 1455–1468, Mar. 1999.

[19] S. Lee, A. Potamianos, and S. Narayanan, "Analysis of Children's Speech: Duration, Pitch and Formants", in *Proc. European Conf. on Speech Communications and Technology*, (Rhodes, Greece), Sept. 1997.

[20] H. Levin and I. Silverman, "Hesitation phenomena in children's speech," *Language and Speech*, vol. 8, pp. 67–85, 1965.

[21] P. Martland, S. P. Whiteside, S. W. Beet, and L. Baghai-Ravary, "Estimating child and adolescent formant frequency values from adult data," in *Internat. Conf. Speech Language Processing*, (Philadelphia, PA), pp. 626–630, Oct. 1996.

[22] J. Mostow, A. G. Hauptmann, and S. F. Roth, "Demonstration of a reading coach that listens," *Proceedings of the ACM Symposium on User Interface Software and Technology*, pp. 77–78, 1995.

[23] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 65–78, Feb. 2002.

[24] A. Potamianos and S. Narayanan, "Spoken dialog systems for children," in *Proc. Internat. Conf. on Acoust., Speech, and Signal Process.*, (Seattle,Washington), pp. 197–201, May 1998.

[25] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 603–616, Nov. 2003.

[26] M. Russell, B. Brown, A. Skilling, R. Series, J. Wallace, B. Bonham, and P. Barker, "Applications of automatic speech recognition to speech and language development in young children," in *Internat. Conf. Speech Language Processing*, (Philadelphia, PA), Oct. 1996.

[27] E. F. Strommen and F. S. Frome, "Talking back to big bird: Preschool users and a simple speech recognition system," *Educational Technology Research and Development*, vol. 41, pp. 5–16, 1993.

[28] B. Xiao, C. Girand, and S.L. Oviatt, "Multimodal Integration Patterns in Children," in *Proc. of the 7th Intern. Conf. on Spoken Language Proc.*, 2002.