A Review of ASR Technologies for Children's Speech

Matteo Gerosa[1], Diego Giuliani[1], Shrikanth Narayanan[2], Alexandros Potamianos[3]

[1]FBK Fondazione Bruno Kessler, Povo (TN), Italy

[2]SAIL Lab, Viterbi School of Engineering, University of Southern California, Los Angeles, CA

[3]Dept. of Electronics and Computer Engineering, Tech. Univ. of Crete, Chania, Greece

{gerosa,giuliani}@fbk.eu potam@telecom.tur.gr shri@sipi.usc.edu

ABSTRACT

In this paper, we review: (1) the acoustic and linguistic properties of children's speech for both read and spontaneous speech, and (2) the performance of automatic speech recognition for children with application to spoken dialogue and multimodal dialogue system design. 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 presented. We conclude with a review of relevant applications of spoken dialogue technologies for children.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Speech recognition and synthesis*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Natural language*.

General Terms

Languages, Human Factor, Design

Keywords

Children's speech analysis, Children's speech recognition, Spoken dialogue

1. INTRODUCTION

In recent years, significant progress has been achieved in the field of automatic speech recognition (ASR) and effective spoken dialogue systems have been built and deployed for a number of applications. Most of this research effort, however, has been devoted to developing systems targeting adult speakers. Following the first studies that raised attention to the poor performance of speech recognition systems for children users, increasing attention has been paid to the area of robust speech recognition technologies for children's speech in a variety of application scenarios such as reading tutors, foreign language learning and multimodal human-machine interaction systems.

Developmental changes in speech production introduce age-dependent spectral and temporal variabilities in 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, reading aids and pronunciation tutoring [55, 45, 52]. Recently a number of systems have been implemented with advanced spoken dialogue interfaces and/or multimodal interaction capabilities [46, 33, 11, 12]. Data collected from these systems as well as new available corpora [7, 58, 8] 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 [40], however, corroborating evidence from the literature is also 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. This paper is intended to provide the reader with a concise overview of the challenges posed by children's speech in automatic speech recognition and spoken dialogue design and reviews the solutions investigated for the different system components. Furthermore, the paper provides suggestions for future research directions in the field.

The rest of this paper is organized as follows. Section 2 presents an overview of speech and multimodal corpora used for carrying out relevant studies in the field. Section 3 attempts to summarize fundamental results of analysis of acoustic and linguistic characteristics of children's speech and relates them with results achieved on adult speech. Effects of developmental changes are also discussed together with their implications on acoustic modeling, language modeling, and design of spoken dialogue systems. Developments in acoustic and language modeling for ASR of children's speech and in design of spoken dialogue systems are reviewed in Section 4, while Section 5 reviews relevant applications where ASR technologies have been applied. Final remarks and suggestions for potential research directions are given at the end of the paper in Section 6.

2. CHILDREN CORPORA

Most of the databases of children recordings focus on the 6-18 age group (or a subset thereof) where collection con-

ICMI-MLMI'09 Workshop on Child, Computer and Interaction November 5, 2009, Cambridge, MA, USA

ditions 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 [40], the KIDS corpus [23], the CU Kids' Audio Speech Corpus [33] and the PF-STAR corpus available in the following languages: British English, Italian, German and Swedish [7].

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 [11], the NICE fairy-tale corpus is presented, where children use open-ended spoken dialogue to interact with animated characters in a game setting. In [8], a child-robot interaction corpus is presented; children interacted with an AIBO robot in open-ended scenarios. In [46], a corpus collected in a Wizard-of-Oz scenario, where children used speech to play a computer game and interact with animated characters on screen is presented and analyzed. In [58], a corpus of child-machine interaction via a multimodal voice and pen interface was collected and analyzed.

For the younger age group (up to 6 year old) there are significant resources thanks to the efforts of the language acquisition community. Most of the available corpora are available via CHILDES, the child language component of the TalkBank system used "for sharing and studying conversational interactions". CHILDES contains over 100 different corpora [42]. Recent data collection efforts include the daily audio-visual recordings of children in their home environment (SpeechHome project) and the weekly audio-only longitudinal recording using the LENA device (Infoture database). Unfortunately, neither of these databases are publicly available.

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 children ages 2-6.

3. SPEECH ANALYSIS

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 [21], and later summarized by Kent [38], 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* [21, 38, 40]. Parametric models for transforming vowel formant frequency of children speakers to the adult speaker space (vowel formant frequency normalization) were considered in [31, 44, 49]. Similarly, a detailed comparison of temporal features and speech segment durations for children and adult speakers can be found in [39, 40]. Again, distinct age-related differences 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 [40]. 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 [28].

3.1 Age trends of acoustic correlates

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 [40]. Recent work on the analysis of the acoustic characteristic of children speech in other languages provided similar results, e.g., see [26] 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 with 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. 1(c),(d) 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 [21] and children speech contains more disfluencies and extraneous speech [55]. An important point is that such results are highly dependent on whether the data was read or spontaneous speech.

3.2 Linguistic Analysis

Some insights regarding the acoustic and linguistic characteristics of children's spontaneous speech can be obtained from the results in [48, 24]. The analysis is based on data from a Wizard-of-Oz study using 160 children playing a voice-activated computer game (Carmen Sandiego corpus [46]). Results show significant age and gender trends.

As far as duration and speaking rate metrics are concerned there is a significant difference between the results for read speech reported in [40] and spontaneous speech reported

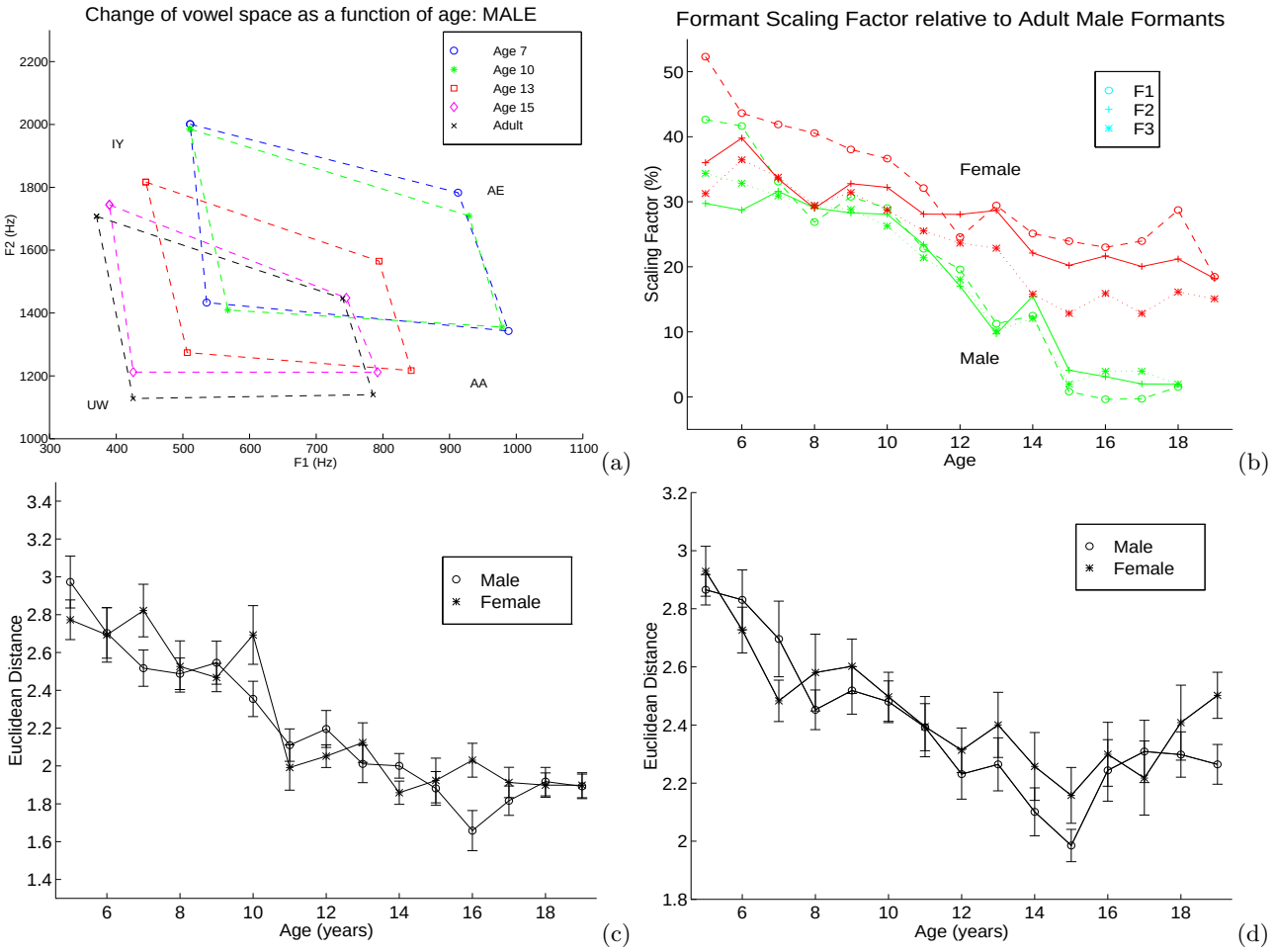

Figure 1: (a) Changes in F1-F2 vowel space as a function of age. The vowel space boundaries are marked by average formant frequency values for the four point vowels /AA, IY, UW, AE/ for the age groups: 7, 10, 13, 15 and adults. (b) Scaling factor variation in first three formant frequencies with respect to age for male and female children (scaling with respect to adult males). Intra-speaker variability as a function of age: (c) Mean cepstral distance between the two repetitions of the same vowels and (d) Mean cepstral distance between the first– and second–half segments within the same vowel realization (from [49]).

here. Specifically, vowel durations are significantly lower and speaking rate is higher for spontaneous than for read speech. The age trend (reduction in duration, increase in speaking rate) was similar for read and spontaneous speech, but adult-level values were reached 1-2 years earlier for spontaneous speech. In general, these differences in duration could be attributed to the cognitive load incurred by the reading task.

In general, the ability of the children to use language efficiently to achieve a task improves with age. Disfluencies decrease with age and children reach adult-skill level at around 12-13 years of age (somewhat earlier for boys than girls. The age trend is reversed for hesitations. The high number of hesitations for girls aged 10-11 compared to boys of the same age group is hard to interpret and could be due to social reasons (rather than linguistic skill). Children use less words per utterance to convey the same message, and, in general, use linguistically simpler constructs as they become more adept with using language over the years. Linguistic variability is reduced with age and older children keep repeating linguistic patterns that have been successful at achieving a specific task. It is interesting to note that for girls in the 12-14 age-group the linguistic variability increases as does the average sentence length. In fact, sentence length increases also for girls aged 10-11 compared to the 8-9 age group. In general, *girls show more linguistic exploration* than boys in the 12-14 age group. This trend seems to emerge around 11 years of age. It is unclear if this trend also correlates with the fact that the specific game is "easy" for older children, i.e., for girls aged 12-14 the game is no longer challenging and thus the opportunity emerges to explore more complex and interesting linguistic patterns. One might conclude that girls ages 12 and older consider language as part of the game not just a tool to successfully complete the game.

In [11], 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 [58], 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.

4. ASR TECHNOLOGIES

As shown in the previous section, acoustic and linguistic characteristics of children's speech are widely different from those of adults' speech and vary rapidly as a function of age. As a consequence, automatic speech recognition systems trained on adults' speech tend to perform poorly on children's speech [57, 50, 15]. However, even in case of a system trained on adequate amount of age-specific speech data, recognition performance reported for children is usually significantly lower than that reported for adults on the same task and it improves as the children's age increases [57, 50, 33, 22]. This correlates well with results of experiments of human perception of speech from children aged 6-11 which have shown that the human word recognition error rate increases as the age of the child decreases [18]. Main factors that concur to make recognition of children's speech more difficult than recognition of adult speech have been analyzed in the previous section.

ASR systems for children's speech widely borrow architectural choices, approaches and algorithms from state-of-the-art ASR systems developed to recognize adult speech. For example, speech signal is often parametrized by mel frequency cepstral coefficients (MFCCs) [35], speech units are almost always modeled by (context-dependent) hidden Markov models (HMMs) [35], and the language model is usually represented by statistical language models based on n-gram statistics [35]. Over the years, a better understanding of characteristics of children's speech has allowed to design and implement solutions suitable for recognizing speech from children of different ages. This section aims at reviewing most relevant developments in automatic recognition of children's speech.

4.1 Acoustic modeling

The influence of a speaker's age on the recognition accuracy of a speech recognizer was initially investigated in [57] for a connected digit recognition task with telephone speech. It was shown that the error rate of a speech recognizer trained with data from speakers of all ages, that is adults, children and elderly speakers, increased significantly for speakers which were twelve years old or younger. Furthermore, recognition performance for children was much lower than that achieved for adults even when using a recognition system trained on children's speech. Correlation of recognition performance with children's age was reported in several papers [57, 50, 33, 22]. In particular, in [33] a huge American English corpus, consisting of speech data collected from children in grades K (kindergarten) through 5, was used to train a specific set of acoustic models for each grade. For test speakers in each grade, recognition results achieved by using the corresponding grade-dependent acoustic models were reported for an isolated word recognition task. It was shown that even using grade-dependent models recognition performance decreased with age and word error rate achieved for children in grade K was more than 100% higher than that achieved for children in grade 5.

Despite that many works confirming the acoustic difference between adult and children's speech, the relative scarcity of large, publicly-available corpora of children's speech induced researchers to study the possibility to employ speech recognizers trained on adult speech to decode children's speech. Earlier works focused on compensating the acoustic variations induced by difference in vocal tract length, which is one of the major source of acoustic variation between children's and adult speech, by using vocal tract length normalization (VTLN) [5]. VTLN is a speaker normalization method that aims at reducing inter-speaker acoustic variability due to different vocal tract length by warping the frequency axis of the speech spectrum of each speaker. In [51] a strong relationship between the optimal warping factor and the age of the speakers was shown when the warping factor selection is performed with respect to HMMs trained on adult speech. Investigations on VTLN [13, 50, 15, 20, 54, 29, 22, 17, 26] show that when a speech recognizer trained on adult speech is applied to decode children's speech, VTLN is able to significantly improve recognition performance. However, recognition results achieved with this compensation approach are still sub-optimal as there are other factors, in addition to differences in vocal tract length, that concur to make adult speech different from children's speech. General acoustic model adaptation techniques such maximum *a posteriori* and maximum likelihood linear regression adaptation can be used to further improve recognition performance [22, 26].

To ensure better recognition performance, age-dependent acoustic models (AMs), trained on speech collected from children, are commonly employed [57, 50, 15, 20, 29]. In principle, training specific models for each target age, or age group, should lead to best performance [57, 33, 19, 22, 14]. However, in order to reduce the amount of data to be collected for robustly training acoustic models, children are often treated as a homogeneous population group and acoustic models are trained with speech from children of all ages [50, 20, 29]. When training age-dependent acoustic models, VTLN is often adopted to further reduce inter-speaker variability [29, 33]. Generally, popular speaker adaptive acoustic modeling methods commonly adopted to train large vocabulary continuous speech recognition systems for adults were also shown effective to train age-dependent AMs for children [34, 30].

Age-dependent AMs, for a target age or age group, and speaker adaptive acoustic modeling represent popular approaches for building AMs to be used in ASR applications. For example, in [34] a reading tutor addressing elementary school children is presented. AMs are trained on children speech by employing speaker adaptive training and VTLN. With a medium size vocabulary, about 1000-2000 words, recognition results achieved on read speech are in the order of 10% word error rate (WER). A noticeable application which makes use of large vocabulary speech recognition for children is presented in [47], where a speech-oriented guidance system with adult and child discrimination capabilities is described. This system has a recognition vocabulary size of 40k words and makes use of different acoustic and language models for adults and children. Thus, even if the system addresses users of all ages, the models used for recognition are still age-dependent and the system relies on its discriminative capability to use the best models to deal with a particular user.

Lately a different approach was proposed in [25, 27] by considering adults and children, in the age range 7-13, as a single population of speakers. Age-independent acoustic models were first conventionally trained by exploiting a small amount (9 hours) of children's speech and a more significant amount (57 hours) of adult speech, for a total of 66 hours of speech. When these age-independent models are used to recognize adult and children's speech, performance decreases for both adults and children compared to the use of age-dependent AMs, trained separately on adults and children's data. Using speaker adaptive acoustic modeling techniques when training on the unbalanced mixture of adults and children's data ensured recognition performance, for both adults and children, as good as that achieved with age-

dependent models. Additional experiments with a recognition vocabulary of 64k words and a trigram language model were carried out on two parallel corpora consisting of the same sentences read by adults and children in the age range 8-12. The WER achieved for children, 10.2%, was only 24% (relative) higher than the WER achieved for adults, 8.2%, thus demonstrating that for the age-range considered, large vocabulary recognition of read children's speech is a feasible task.

Additional areas of potential improvement of recognition performance for children's are represented by acoustic feature extraction, speech pattern duration modeling and pronunciation modeling [57, 50, 41]. The effect of frequency bandwidth reduction on recognition of children's speech was investigated in [20, 41]. In particular, in [41], the sampling rate for children's speech was downsampled from the original 20kHz to 2kHz. Similarly was done with adult speech, but starting from the original sampling rate of 16kHz. HMM training and recognition were repeated for each sampling rate. For children, the decrease of performance was relatively small down to 6kHz. A significant degradation in performance was observed between 4kHz and 2kHz for both children and adults, but degradation was much greater for children. It was observed that most values of the third formant for children's speech fall outside telephone bandwidth, and this could explain well the low recognition performance reported for telephone applications with children [57].

As mentioned above, the acoustic front-end of an ASR system for children is often based on standard MFCCs. Attempts to find out better acoustic features for children's speech did not succeed. For example, in [57] the effectiveness of LPC-based cepstral parameters and mel based cepstral features were compared in the context of a connected digit recognition task with telephone speech. The use of mel based cepstral features resulted in better recognition performance. In [56] a special variation in the mel filter bank was investigated, consisting of the normalization of the spectral envelopes using a technique called weighted overlapped spectral averaging (WOSA). Using this front-end it was shown that it is more appropriate to assume that the spectral envelopes of any two speakers are scaled version of one another rather than whole magnitude spectrum including pitch harmonics.

Idealized baseforms in the pronunciation dictionary of an ASR system may result unsuitable for children with a poor pronunciation or for younger children. The use of a user customized pronunciation dictionary was investigated in [41]. It was found that a simple modification of the pronunciation dictionary improves ASR perforce only to a limited extent. In [14] the use of a customized pronunciation dictionary for a specific target age group was investigated. A specific pronunciation dictionary was developed for preschool children deriving new pronunciation rules by looking into training data on how words were actually pronounced by preschool children with respect to the expected standard pronunciation. Based on these rules, extra pronunciation variants were then added into the standard pronunciation dictionary. Significant performance improvements were observed in comparison to when using the standard pronunciation dictionary, however these improvements were found vanishing when acoustic models were trained mainly with speech data from preschool children. This is because pronunciation variations, specific of preschool children, were already "learned" by the acoustic models during training exploiting the standard pronunciation dictionary.

Despite many works confirming the difference in speech pattern duration between adult and children's speech and between children of different ages (see the previous section), duration modeling in ASR of children's speech is a topic not yet investigated.

4.2 Spoken Dialogue

Voice interaction as a component of the multimedia experience fabric is an input modality greatly desired by children users. The addition of conversational capability to children's multimedia applications contributes to more natural user interaction and improved user experience. Although higher variability and different interaction patterns create additional challenges, there has been notable efforts in the literature for designing, implementing and testing prototype multimodal systems for children. Early spoken dialogue application prototypes that were specifically aimed at children included word games for pre-schoolers [55], aids for reading [45] and pronunciation tutoring [52]. Recently a number of systems have been implemented with advanced spoken dialogue interfaces, multimodal interaction capabilities and/or embodied conversational characters [46, 33, 11, 12]. A well-known dialogue system for children was developed in the context of the NICE project [**?**]. In this project users of all ages interact with lifelike conversational characters in a fairy-tale world inspired by the Danish author H.C. Andersen. Almost all of the aforementioned systems have focused in the age group 6-15. In [36], a task oriented multimodal dialogue system for preschoolers with fantasy and curiosity elements was implemented and evaluated. It was shown that speech interaction was a motivator for young children to finish the tasks at hand.

Building successful spoken dialogue systems requires both good technology and good interface design. In addition to building acoustic models customized to children as outlined above, pronounciation and language modeling is also an issue. The importance of using customized language models in recognition of spontaneous children's speech has been pointed out in several works [20, 46, 14]. Effectiveness of training the language model by exploiting task related written text or manual transcriptions of utterances collected from past users of the system has been shown.

Experience for building and evaluating multimodal dialogue systems for children shows that emphasis has to be placed on interface design. For example, it has been found that using animated sequences to communicate information and adding 'personality' to the interface significantly improved the user experience. In addition, the flexible choice of input modality (any of speech, natural language, commands or buttons) make the application easy to use even for novice users or users that are not adept at using a specific modality, e.g., pre-school children are not efficient users of keyboard/mouse.

For children (especially young children) learning and playing are intertwined activities. Thus the main goal of a successful dialogue system is to provide fun, excitement and engagement. In [36], it is shown that fantasy and curiosity elements in task-oriented dialogue systems for preschoolers increase user satisfaction and in some cases also user engagement. In general, balancing exploration (e.g., open-ended games, story telling) and exploitation (e.g., task-oriented games, arcade games) in spoken dialogue and multimodal systems for children is an open research issue that requires more attention. Interface and application design recommendations for children are expected to be both age- and gender-specific, e.g., it is well known that the attention span for children grows with age and that girls prefer games with a strong social interaction aspect.

4.3 Emotion

An essential step toward building natural and respon-

sive spoken interaction systems, especially for children, is to analyze and detect age- and gender-dependent user behaviors. In [4], polite and frustrated behavior of children during spoken dialogue interaction with computer characters in a computer game was analyzed. Results were consistent with research results from language acquisition showing that even six and seven year-old children have awareness and command of varying levels of politeness [4]. It has also been shown that children use impoliteness (insult) more frequently than adults when interacting with spoken dialogue system [10].

The analysis in [6] showed that children aged 10-11 were significantly more polite and less frustrated than older and younger children. As far as gender is concerned, girls are significantly more polite and less often frustrated than boys. The frustration age trend can be partially attributed to the game challenge factor and task completion, e.g., verbal expressions of frustration occurred more than twice as often in games that ended up in a loss than in those that were won [6]. Frustration went up significantly when recognition errors were involved (similar results are reported for adults [32]). Recognition errors seem to be irritating certain children much more than others. The age trend in the level of politeness was partially explained by the "social standing" that the child attributes to the animated characters, as well as, the challenge that the game provides.

Overall, girls are more polite and are less often frustrated than boys in spoken child-computer interaction. In addition, child age, social roles, and game design significantly affects children's choices about politeness and frustration. Analysis also showed that some common "warning words" were especially salient in indicating polite and frustrated behavior. In addition to lexical markers, pragmatic markers, e.g., repetition, were good indicators of frustration.

As far as emotion recognition is concerned, results in [59] show that lexical cues have more discriminative power than acoustic and dialogue cues for detection of politeness, whereas dialogue and acoustic cues are better for frustration detection. This is in agreement with the analysis results that show that politeness is more explicitly marked in language usage, while repetitions and corrections (due to system errors or task difficulty) lead to frustration. Based on the results of both two-way and three-way classification experiments it is clear that by augmenting acoustic features with lexical and contextual information classification performance improves significantly. The results also showed age and gender trends, e.g., classification performance was better for girls than for boys.

Another comprehensive study on emotion recognition, focusing on mono-modal systems with speech as only input channel, is reported in [53]. The main results are achieved on the FAU Aibo Emotion Corpus, a corpus of spontaneous, emotionally colored speech of children at the age of 10 to 13 years interacting with the Sony robot Aibo. In this work a classwise averaged recognition rate of almost 70% for the 4-class problem 'Anger', 'Emphatic', 'Neutral', and 'Motherese', has been reported by combining both acoustic and linguistic features.

Initial results for emotion recognition of preschoolers interacting is a task-oriented multimodal dialogue system can be found in [37].

5. APPLICATIONS

As soon as the ASR technology started to become more robust, it was exploited for building prototype systems for a variety of applications. Some relevant examples are briefly described below.

Language learning and assessment.

Assessment of a child's reading abilities requires personal attention by teachers, which is a time-consuming process that decreases actual time spent on instruction. Several ASR-based applications have been developed to partially alleviate teachers' workload by assessing students individually. Following first attempts to use ASR technologies in automated pronunciation [52] and reading [45] tutors, several system prototypes have been developed and assessed in real operating environments. Some noticeable examples are listed below.

In the LISTEN project [9], CMU researchers are developing an ASR-based tutor that listens to children read aloud and analyzes student's oral reading (grades 1-5.) The University of Colorado's *Foundations to Literacy* program is a comprehensive reading program designed for beginning and poor readers [34]. It consists of a set of tightly integrated computer-based multi-modal learning applications in which children interact with a Virtual Tutor, Marni, to learn to read well. The TBALL project [3] concerns the design and realization of an automatic system for assessing and evaluating the language and literacy skills of young children. The system aims at automatically assessing the English literacy skills of mainly Mexican-American children in grades K-2, and is composed by several assessment modules that make use of ASR-based technology.

Second and foreign language learning is also an area in which ASR technology has found a natural application especially for speech-recognition-based pronunciation training. In this area of application there are already several commercialized systems targetting children [1, 2].

Diagnosis and remedial treatments of pathological speech.

A variety of ASR technology enabled applications for children with special needs have been addressed. Two noteworthy system prototypes are described below.

In [16], the experience acquired using Baldi, an animated conversational agent, in daily classroom activities with profoundly deaf children is presented. One component of conversing with Baldi is automatic recognition of a deaf child's speech. This is an extremely challenging area for ASR which involves recognition of deaf children's speech and properly rejection of incorrectly-pronounced words.

PEAKS is an automatic assessment system of the intelligibility of speech which can be accessed via the internet [43]. This tool allows, for example, to assess speech from children with cleft lip and palate (CLP). CLP is one of the most common alterations of the face. Speech of children with CLP shows particular characteristics, which can result in speech disorders also after surgical treatment. In this context, the diagnosis of speech disorders is of crucial importance for improving, and monitoring over time, speech intelligibility.

Toys and games.

The toy industry has been moving towards the target of interactive toys in the past two decades. Although speech plays an increasingly important role in toys, speech interaction is often one way, i.e., toys speak prerecorded prompts activated by pressure sensitive sensors. Starting from TI's Speak and Spell, there has been a variety of toys that incorporate speech recognition or speech identification capabilities, e.g., toys with voice passwords. In the past decade, robotic toys and interactive pets with the ability to understand a limited vocabulary have emerged, e.g., Sony AIBO, Mattel's Diva Petz, MGA's Commando-Bot. Recenlty, robotic toys are becoming increasingly interactive and able to understand complex commands.

In the past two decades, there have been numerous efforts to incorporate speech recognition technology in desktop games. In most cases, these efforts had limited success, as game developers considered speech as just another input modality, rather than re-designing the application and interface around the speech modality. A good example from the research world of how to design multimodal dialogue games is the NICE project, where an open-ended story-telling environment was enhanced by incorporating interactive animated characters [**?**]. In the past few years, speech recognition technology has found its way into all major console gaming platforms. Development of successful spoken dialogue games should follow, provided that game developers will do away with the notion of speech as just another command and control modality, and embrace the conversational, social and interactive aspects of speech interaction.

6. CONCLUSIONS

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).

7. REFERENCES

[1] Aurolog, Phoenix, AZ, USA. http://www.tellmemore.com.

[2] Rosetta Stone Ltd., Arlington, VA, USA. http://www.rosettastone.com.

[3] A. Alwan, Y. Bai, M. Black, L. Casey, M. Gerosa, M. Heritage, M. Iseli, B. Jones, A. Kazemzadeh, S. Lee, S. Narayanan, P. Price, J. Tepperman, and S. Wang. A System for Technology Based Assessment of Language and Literacy in Young Children: the Role of Multiple Information Sources. In *Proc. of International Workshop on Multimedia Signal Processing*, Chania, Creete, GREECE, Oct 2007.

[4] E. Andersen, M. Brizuela, B. DuPuy, and L. Gonnerman. Cross-linguistic evidence for the early acquistion of discourse markers as register variable. *Journal of Pragmatics*, (31):1339–1351, 1999.

[5] A. Andreaou, T. Kamm, and J. Cohen. Experiments in Vocal Tract Length Normalization. In *Proc. of the CAIP Workshop: Frontiers in Speech Recognition*, 1994.

[6] S. Arunachalam, D. Gould, E. Andersen, D. Byrd, and S. Narayanan. Politness and frustration language in child-computer interactions. In *Proceedings of Eurospeech*, pages 2675–2678, Aalborg, Denmark, 2001.

[7] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong. The PF-STAR Children's Speech Corpus. In *Proc. of INTERSPEECH/ICSLP*, pages 2761–2764, Lisboa, Portugal, Sep. 2005.

[8] 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.

[9] J. Beck, P. Jia, and J. Mostow. Automatically assessing oral reading fluency in a computer tutor that listens. In *Technology, Instruction, Cognition and Learning*, volume 1, pages 61–81, 2004.

[10] L. Bell. *Linguistic Adaptations in Spoken Human-Computer Dialogues - Empirical Studies of User Behavior*. PhD thesis, KTH, Sweden, 2003.

[11] 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. Interspeech*, pages 2765–2768, Lisbon, Portugal, 2005.

[12] L. Bell and J. Gustafson. Children's convergence in referring expressions to graphical objects in a speech-enabled computer game. In *Proc. Interspeech*, pages 2209–2212, Antwerp, Belgium, 2007.

[13] D. C. Burnett and M. Fanty. Rapid Unsupervised Adaptation to Children's Speech on a Connected-Digit Task. In *Proc. of ICSLP*, volume 2, pages 1145–1148, Philadelphia, PA, 1996.

[14] T. Cincarek, I. Shindo, T. Toda, H. Saruwatari, and K. Shikano. Development of Preschool Children Subsystem for ASR and Q&A in a Real-Environment Speech-Oriented Guidance Task. In *Proc. of INTERSPEECH/ICSLP*, pages 1469–1472, 2007.

[15] T. Claes, I. Dologlou, L. ten Bosch, and D. V. Compernolle. A Novel Feature Transformation for Vocal Tract Length Normalisation in Automatic Speech Recognition. *IEEE Trans. on Speech and Audio Processing*, 6(6):549–557, Nov. 1998.

[16] R. Cole, D. Massaro, B. Rundle, K. Shobaki, J. Wouters, M. Cohen, J. Beskow, P. Stone, P. Connors, A. Tarachow, and D. Solcher. New Tools for Interactive Speech and Language Training: Using Animated Conversational Agents in the Classrooms of Profoundly Deaf Children . In *Proceedings of ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education*, Apr. 1999.

[17] X. Cui and A. Alwan. Adaptation of children's speech with limited data based on formant-like peak alignment. *Computer Speech and Language*, 20:400–419, Jul. 2006.

[18] S. D'Arcy and M. Russell. A Comparison of Human and Computer Recognition Accuracy for Children's Speech. In *Proc. of INTERSPEECH/ICSLP*, pages 2197–2199, Lisboa, Portugal, Sept. 2005.

[19] S. D'arcy, L. P. Wong, and M. Russel. Recognition of read and spontaneous children's speech using two new corpora. In *Proc. of INTERSPEECH/ICSLP*, pages 1473–1476, Jeju Island, Korea, Oct. 2004.

[20] S. Das, D. Nix, and M. Picheny. Improvements in Children's Speech Recognition Performance. In *Proc. of ICASSP*, pages 433–436, Seattle,WA, May 1998.

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

[22] D. Elenius and M. Blomberg. Adaptation and Normalization Experiments in Speech Recognition for 4 to 8 Year Old Children. In *Proc. of INTERSPEECH*, pages 2749–2752, Lisbon, Portugal, Sept. 2005.

[23] M. Eskernazi. Kids: A database of children's speech. *Journal of the Acoustical Society of America*, 100(4):2759–2759, 1996.

[24] V. Farantouri, A. Potamianos, and S. Narayanan. Linguistic analysis of spontaneous children speech. In *Workshop on Child, Computer and Interaction*, Chania, Greece, 2008.

[25] M. Gerosa, D. Giuliani, and F. Brugnara. Speaker Adaptive Acoustic Modeling with Mixture of Adult

and Children's Speech. In *Proc. of INTERSPEECH/ICSLP*, pages 2193–2196, Lisboa, Portugal, Sep. 2005.

[26] M. Gerosa, D. Giuliani, and F. Brugnara. Acoustic variability and automatic recognition of children's speech. *Speech Communication*, 49:847–869, 2007.

[27] M. Gerosa, D. Giuliani, and F. Brugnara. Towards age-independent acoustic modeling. *Speech Communication*, 51:499–509, 2009.

[28] M. Gerosa, S. Lee, D. Giuliani, and S. Narayanan. Analyzing Children's Speech: An acoustic Study of Consonants and Consonant-Vowel Transition. In *Proc. of ICASSP*, pages 393–396, Toulouse, France, May 2006.

[29] D. Giuliani and M. Gerosa. Investigating Recognition of Children Speech. In *Proc. of ICASSP*, volume 2, pages 137–140, Hong Kong, Apr. 2003.

[30] D. Giuliani, M. Gerosa, and F. Brugnara. Improved automatic speech recognition through speaker normalization. *Computer Speech and Language*, 20(1):107–123, Jan. 2006.

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

[32] J. Gustafson and L. Bell. Speech technology on trial - experiences from the august system. *Natural Language Engineering*, 6(3-4):273–286, 2000.

[33] A. Hagen, B. Pellom, and R. Cole. Children's Speech Recognition with Application to Interactive Books and Tutors. In *Proc. of IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, pages 186– 191, St. Thomas, US Virgin Islands, Dec. 2003.

[34] A. Hagen, B. Pellom, S. V. Vuuren, and R. Cole. Advances in Children's Speech Recognition within an Interactive Literacy Tutor. In *Proc. of HLT/NAACL*, pages 25–28, Boston, MA, May 2004.

[35] X. Huang, A. Acero, and H.-W. Hon. *"Spoken Language Processing"*. Prentice Hall, New Jersey, 2001.

[36] T. Kannetis and A. Potamianos. Towards adapting fantasy, curiosity and challenge in multimodal dialogue systems for preschoolers. In *Internat. Conf. on Multimodal Interfaces*, Cambridge, MA, Nov. 2009.

[37] T. Kannetis, A. Potamianos, and G. Yannakakis. Fantasy, curiosity and challenge as adaptation indicators in multimodal dialogue systems for preschoolers. In *Workshop on Child, Computer and Interaction*, Cambridge, MA, Nov. 2009.

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

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

[40] 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*, pages 1455–1468, Mar. 1999.

[41] Q. Li and M. Russell. An Analysis of the Causes of Increased Error Rates in Children's Speech Recognition. In *Proc. of ICSLP*, pages 2337–2340, Denver, CO, Sep. 2002.

[42] B. MacWhinney. *The CHILDES Project: Tools for Analyzing Talk (3rd Ed.)*. Lawrence Erlbaum Associates, Mahwah, NJ, 2000.

[43] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth. PEAKS - A system for the automatic evaluation of voice and speech disorders. *Speech Communication*, 51(5):425–437, 2009.

[44] 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*, pages 626–630, Philadelphia, PA, Oct. 1996.

[45] J. Mostow, S. Roth, A. Hauptmann, and M. Kane. A prototype reading coach that listens. In *"Proc. 12th Natl. Conf. on Artificial Intelligence (AAAI-94)"*, pages 785–792, Seattle, WA, 1994.

[46] S. Narayanan and A. Potamianos. Creating conversational interfaces for children. *IEEE Transactions on Speech and Audio Processing*, 10(2):65–78, Feb. 2002.

[47] R. Nisimura, A. Lee, H. Saruwatari, and K. Shikano. Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability. In *Proc. of ICASSP*, volume 1, pages 433–436, Montreal, Canada, May 2004.

[48] A. Potamianos and S. Narayanan. Spoken dialog systems for children. In *Proc. Internat. Conf. on Acoust., Speech, and Signal Process.*, pages 197–201, Seattle,Washington, May 1998.

[49] A. Potamianos and S. Narayanan. Robust recognition of children's speech. *IEEE Transactions on Speech and Audio Processing*, 11(6):603–616, Nov. 2003.

[50] A. Potamianos, S. Narayanan, and S. Lee. Automatic Speech Recognition for Children. In *Proc. of Eurospeech*, pages 2371–2374, Rhodes, Greece, Sept. 1997.

[51] A. Potamianos and R. Rose. On Combining Frequency Warping and Spectral Shaping in HMM Based Speech Recognition. In *Proc. of ICASSP*, volume 2, pages 1275–1278, Munich, Germany, Apr. 1997.

[52] M. Russell, C. 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 *Proc. of ICSLP*, pages 176–179, Philadelphia, PA, 1996.

[53] S. Steidl. *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. PhD thesis, Friedrich-Alexander University, Erlangen, Germany, 2009.

[54] G. Stemmer, C. Hacker, S. Steidl, and E. Nöth. Acoustic Normalization of Children's Speech. In *Proc. of Eurospeech*, pages 1313–1316, Geneva, Switzerland, Sept. 2003.

[55] 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*, 41:5–16, 1993.

[56] S. Umesh, R. Sinha, and S. V. B. Kumar. An investigation into front-end signal processing for speaker normalization. In *Proc. of ICASSP*, volume 1, pages 345–348, Montreal, Canada, May 2004.

[57] J. Wilpon and C. Jacobsen. A Study of Speech Recognition for Children and Elderly. In *Proc. of ICASSP*, pages I–349–352, Atlanta,GA, May 1996.

[58] B. Xiao, C. Girand, and S. Oviatt. Multimodal integration patterns in children. In *Proc. of the 7th Intern. Conf. on Spoken Language Proc.*, pages 629–632, 2002.

[59] S. Yildirim, C. Lee, S. Lee, A. Potamianos, and S. Narayanan. Detecting politeness and frustration state of a child in a conversational computer game. In *Proc. European Conf. on Speech Communication and Technology*, pages 2209–2212, Lisbon, Portugal, Sept. 2005.