Linguistic Analysis of Spontaneous Children Speech

Vassiliki Farantouri, Alexandros Potamianos and Shrikanth Narayanan†
Dept. of Electronics and Computer Engineering, Tech. Univ. of Crete, Chania 73100, Greece
†Speech Analysis and Interpretation Lab, Univ. of Southern California, CA 90089, USA
{vfarad,potam}@telecom.tuc.gr shri@sipi.usc.edu

ABSTRACT

In this paper, we investigate the duration, lexical and linguistic properties of children's spontaneous speech for children ages 8 to 14 interacting with animated characters in a computer game. Age and gender trends are studied for parameters such as phone and sentence duration, speaking rate, fluency (mispronounciations and hesitations), vocabulary size and linguistic variability measured via bigram language model perplexity. The analysis shows significant differences between read- and spontaneous children speech in terms of absolute values of acoustic and linguistic parameters, as well as, linguistic variability. In addition, spontaneous data present clear gender-specific trends, e.g., increased "language exploration" by girls in the 12-14 age group. The applicability of these results for acoustic and linguistic modeling and spoken dialogue systems interface design is discussed.

1. INTRODUCTION

Children's speech exhibits quite different characteristics compared to adults. In addition, as recent studies have verified, acoustic and linguistic characteristics of children change with age and gender. Studying acoustic and linguistic parameters of children of different age and gender is important for designing and implementing applications suitable for specific children groups. Spontaneous speech is very different from "read speech" on which most of the acoustic analysis studies have been performed in the literature. Recently a number of systems have been implemented with advanced spoken dialogue interfaces and relevant studies have improved our understanding of verbal child-machine interaction. In this paper, we focus on the linguistic characteristics of spontaneous children speech, e.g., vocabulary size, perplexities and linguistic variability for children verbally interacting with computer game animated characters. We also review fluency parameters, e.g., hesitations and mispronunciations, and duration metrics, e.g., speaking rate and sentence duration. We present age/gender trends spotted and the implied age dependency of all these parameters. Important conclusions are drawn about the acoustic and linguistic differences between "read" and spontaneous children speech. In addition, the results clearly demonstrate gender-specific trends that have important implications for application and interface design of children computer interaction systems.

2. PRIOR WORK

Early spoken dialogue application prototypes that were specifically aimed at children included word games for pre-schoolers [26], aids for reading [21] and pronunciation tutoring [25]. Recently a number of systems have been implemented with advanced spoken dialogue interfaces, multimodal interaction capabilities and/or embodied conversational characters [22, 12, 4, 5]. Data collected from these systems as well as new available corpora [2, 27, 3] have improved our understanding of verbal child-machine interaction.

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 ("read speech") that is mostly used for acoustic analysis and modeling are the American English CID children corpus [17], the KIDS corpus [8], the CU Kids' Audio Speech Corpus [12] and the PF-STAR corpus available in the following languages: British English, Italian, German and Swedish [2]. As far as child-machine spontaneous speech interaction is concerned 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 [27], a corpus of child-machine interaction via a multimodal voice and pen interface was collected and analyzed.

The acoustic characteristics of children for "read speech" have been first analyzed in [7, 15] and later on in [17] for American English. Recently such studies have been carried out for other languages as well, e.g., Italian [9]. In all studies, children demonstrate larger fundamental and formant frequency, as well as, *higher acoustic variability.* In general, it is considered that variability converges to adult values around 13-14 years of age [17]. A detailed comparison of temporal features and speech segment durations for children vs adults (for "read speech") can be found in [16, 17]. 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.

There is no detailed analysis in the literature of the acoustic and linguistic characteristics of spontaneous children's speech due to the lack of large corpora. However, there are limited studies of child-machine spontaneous speech interaction using smaller corpora. 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 [1], politeness and frustration markers were analyzed for the CHIMP database (the database also analyzed in this paper). Younger children used politeness markers more commonly and expressed frustration verbally more often than older children. In [27], 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

3. DESCRIPTION OF DATABASE

In this study, we used data collected during the CHIMP (Children's Interactive Multimedia) project [23]. In this project, verbal interaction between children and computer animated character was investigated. The data was collected using a Wizard of oZ (WoZ) experimental setup. The children were verbally interacting with the computer game designed for children ages eight and older. To successfully complete the game, i.e., arrest the appropriate suspect, two subtasks have to be completed, namely, determining the physical characteristics of the suspect to issue an arrest warrant and tracking the suspect's whereabouts. The player talked to animated characters on the game screen and ask them for clues that could be correlated with information in a geographical database. Overall, the game was rich in dialog subtasks including: navigation and multiple queries, database entry, and database search.

Data from a total of 160 children and 7 adult players were collected using the speech WoZ scenario (with no recognition errors). Most players played the game twice. The total number of games played per age group[1] and gender are shown next:

	Age								
Gnd	8	9	10	11	12	13	14	8-14	>21
F	18	23	32	24	10	8	4	119	5
M	21	51	16	23	21	25	14	171	8

A total of about 50000 utterances were collected. In order to obtain statistically significant results, the analysis in this paper focused on three age and gender groups, namely: 8-9, 10-11 and 12-14.

The data was transcribed and annotated for disfluencies and hesitation phenomena. In addition, each utterance was annotated with the emotional state of the child [1]. Child-computer interaction turns were also manually categorized into a set of predefined "dialogue states". Dialog states roughly corresponded to one (or a group of similar) game actions taken by the wizard in response to a voice command. For example, the dialog state "Talk2Him" incorporated voice commands asking for a cartoon character's attention, while states "WhereDid" and "TellMeAbout" correspond to queries about the suspect's whereabouts and physical characteristics, respectively. For more details on the database and data collection procedure see [23, 22, 1]

4. EVALUATION METHODOLOGY

Duration, fluency and lexical Metrics are calculated and presented in this study. Specifically:

- *Duration metrics:* Phone and sentence durations, rate of speech information and between-word silence duration were computed from automatically estimated phone-level corpus segmentation. The segmentation information was obtained using hand-labeled word-level transcriptions and context-independent phone-level hidden Markov models (HMMs) trained on the CHIMP corpus. The HMMs were then used to perform forced segmentation on the transcribed utterances. Note that the analysis was carried out on a subset of the database excluding utterances with disfluencies and hesitations, i.e., outliers.

[1]There was also a limited amount of data collected for 6 and 7 year-olds; these data were excluded from our study.

The following metrics were computed per age (group) and gender:

1. *Phone duration:* Here we focus on vowel duration to be able to directly compare with the average vowel duration results of [17].
2. *Sentence duration.*
3. *Sentence duration per dialogue state.*
4. *Between-word silence duration.*
5. *Speaking rate*: computed as the average number of phones per seconds..
6. *Speaking rate excluding silence*: computed as the average number of phones/sec excluding between word silence segments.

- *Fluency metrics:* False-starts, mispronunciations, hesitations and filled pauses were manually labeled on the spontaneous speech corpus and statistics were computed per age and gender group.

- *Lexical and Syntactic metrics:* Sentence length, vocabulary size, linguistic perplexity and lexical-variation were estimated on the manually transcribed corpus. Perplexities were calculated using bigram language models. Turn-to-turn lexical variability was also measured using the Levenshtein distance. The following metrics were computed:

1. *Sentence Length*: mean number of words and number of phones per sentence computed per age and gender group.
2. *Vocabulary size*: mean number of unique words per speaker, averaged for each age and gender group.
3. *Average number of tokens*: mean number of words per speaker, averaged for each age and gender group.
4. *Linguistic variability*: perplexity of the bigram language model for each age and gender group. Either all data where used for training and perplexity computation (type I) or 2/3 of the data where used for training and the rest for perplexity computation in a round-robin fashion (type II).
5. *Linguistic variability for each dialogue state*: same as above but perplexity (and language models) are computed over each dialogue state separately and then averaged over all dialogue states.
6. *Intra- vs inter- speaker linguistic variability*: computed as the average ratio of the perplexity of the bigram language model of the one speaker's utterances vs all speaker's utterances in an age and gender group.
7. *Linguistic variability turn to turn*: computed as the Levenshtein distance between two adjacent utterances in the *same* dialogue state, for each age and gender group (this metric was also used in [23]). The metric is computed for dialogue states where the user makes the same type of requests, e.g., "WhereDid".

Finally, statistical significance of the results was tested using 2-way ANOVA analysis.

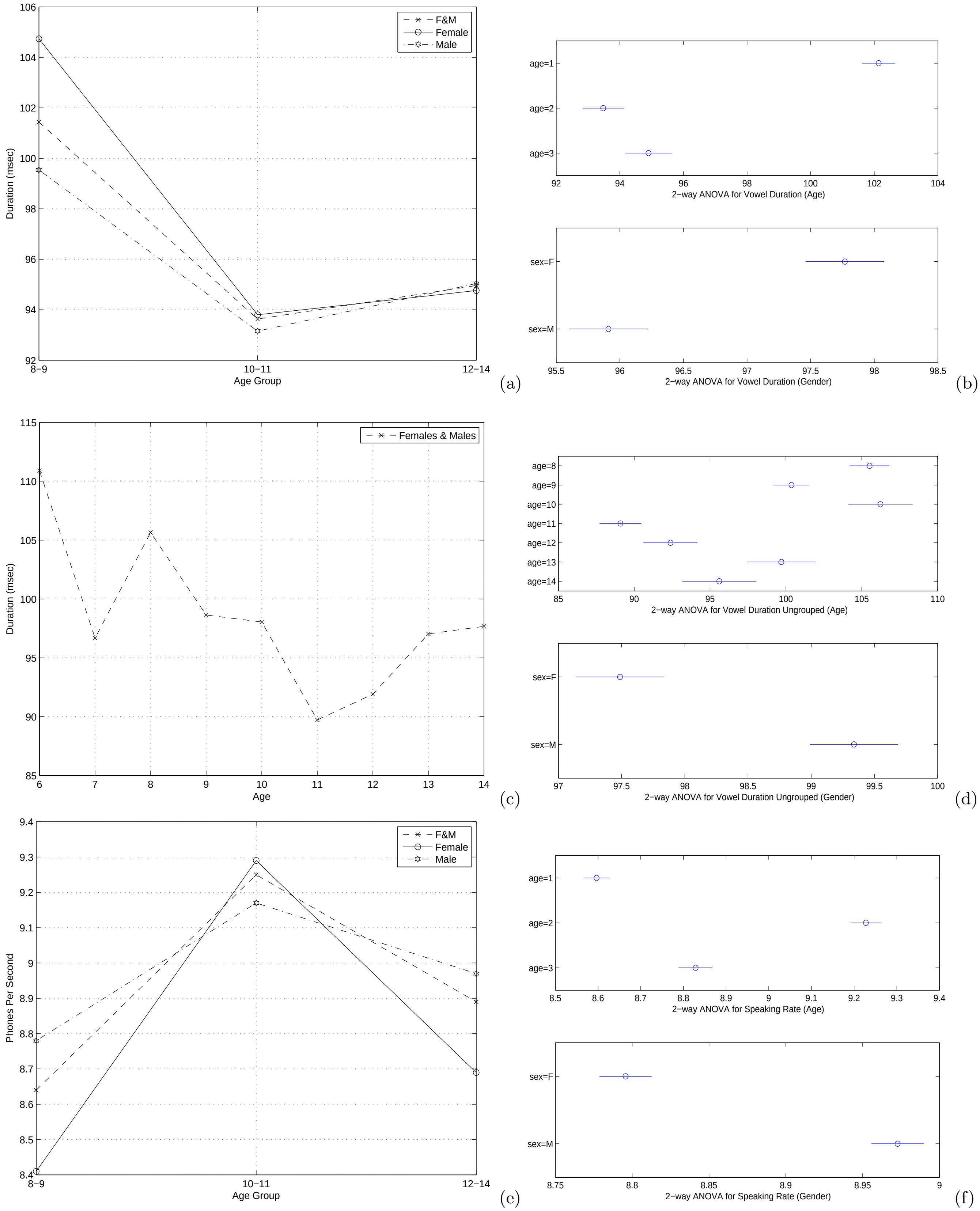

Figure 1: (a) Vowels Duration as a function of age and gender, and (b) 2-way ANOVA results (age-group and gender). (c) Vowels Duration for all ages as a function of age and gender, and (d) 2-way ANOVA results (age-group and gender). (e) Speaking Rate as a function of age and gender, and (f) 2-way ANOVA results (age-group and gender).

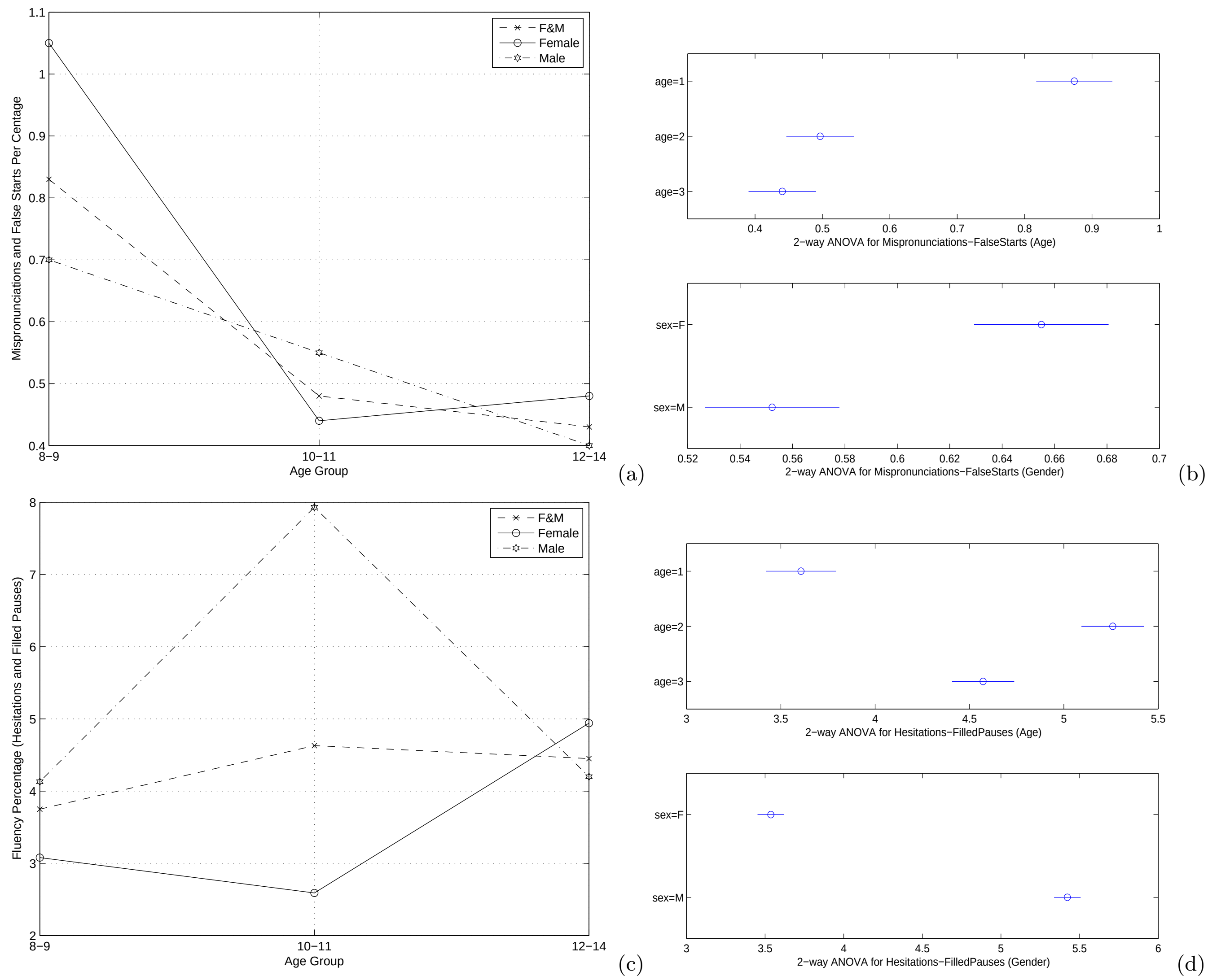

Figure 2: (a) False starts and Mispronunciations Per Word Per Utterance as a function of age and gender, and (b) 2-way ANOVA results (age-group and gender). (c) Hesitations and Filled Pauses Per Words Per Utterance as a function of age and gender, and (d) 2-way ANOVA results (age-group and gender).

5. RESULTS

In this section, we present the most important results for each category namely: duration statistics, fluency statistics, lexical metrics and language model perplexity metrics. Each figure is accompanied with the corresponding 2-way ANOVA analysis with respect to age and gender group. All statistics are computed for three age groups: 8-9, 10-11 and 12-14, with the exception of vowel duration where data are presented (also) for all ages.

In Fig. 1(a), the average vowel phone duration is shown for all age and gender groups. The same information is shown in Fig. 1(c) for all ages[2]. As expected, average vowel duration decreases with age. Specifically, there is a significant reduction in average vowel duration between the younger (8-9) and middle (10-11) age group, and then the duration levels off (and increases somewhat) for the older age group (12-14). The trend is similar for both genders. The same conclusions can be drawn from the more detailed Fig. 1(c) where the average durations decreases for ages 8 to 11 and then increases for ages 12-14.

In Fig. 1(e), the speaking rate is shown in terms of phones per second (including inter-word silences) for all age and gender groups. The middle age group (10-11 years) is speaking significantly faster than the younger and older groups. The differences in speaking rate among the age groups is up to 10% for female speakers. It is interesting to note that there is a significant reduction in speaking rate between the age groups 10-11 and 12-14.

In Fig. 2(a), the average number of false starts and mispronounciations are shown as the percent of the total spoken words. Disfluencies decrease as a function of age. Specifically, there is a significant relative reduction of 30% between the 8-9 and 10-11 age groups. The reduction is even bigger for female speakers. Disfluencies decrease from the 10-11 to the 12-14 age groups, especially for male speakers; however, the reduction is not statistically significant. In Fig. 2(c), the average number of hesitations and filled pauses are shown as a percent of the total spoken words. The trend here is that

[2]Note that the amount of data for ages 6 and 7 is very limited; the data points are only shown for completeness.

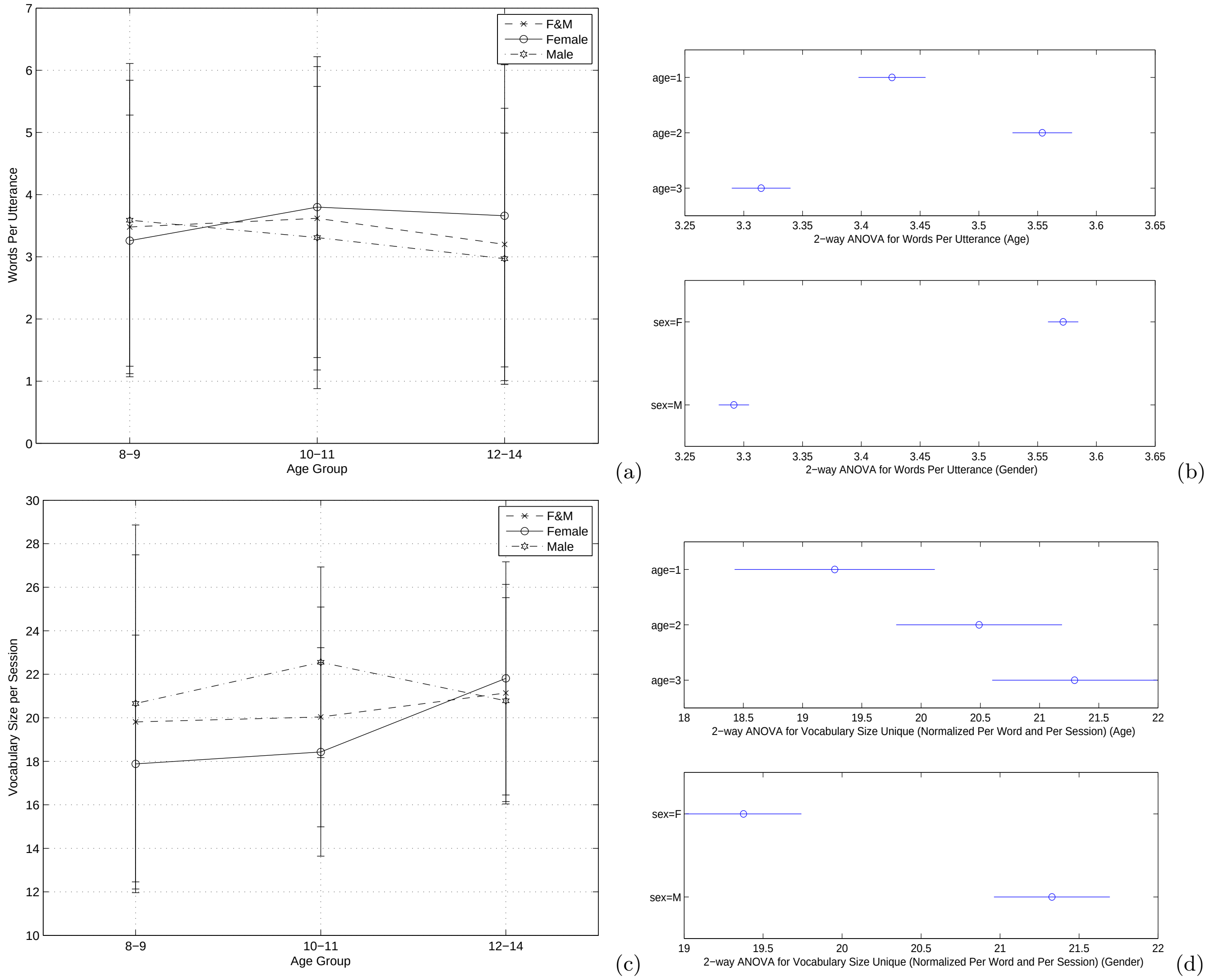

Figure 3: (a) Words Per Utterance as a function of age and gender, and (b) 2-way ANOVA results (age-group and gender). (c) Vocabulary Size Unique Per Words Per Session as a function of age and gender, and (d) 2-way ANOVA results (age-group and gender).

(unlike disfluencies) hesitations increase somewhat with age. This is consistent with the observations in [23]. In addition, there is a clear gender difference for the 10-11 age group, girls tend to hesitate much more than boys (at least twice as much)[3]. When looking at the breakdown of hesitations vs. filled pauses (not shown here) hesitations in the form of breathing noises are significantly higher for younger children, while filled pauses are much more common for older children (see also [23]). 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 12-14 age group.

In Fig. 3(a), the average number of words per utterance is shown as a function of age group and gender. There is clear gender trend here, and girls tend to be more verbose than boys especially for the older age group. For boys, the average number of words per utterance consistently decreases with age, while for girls verbosity increases between the age groups 8-9 and 10-11, and then levels off. The relative difference in verbosity between age groups is between 10 and 20%. In Fig. 3(c), the average vocabulary size per session is shown as a function of age group and gender. Although the average vocabulary size tends to increase with age the age trend is not significant. Surprisingly there is a gender trend: boys have a richer vocabulary than girls for the 8-9 and 10-11 age groups[4].

In Fig. 4, three different measures of linguistic complexity and variability are shown as a function of age group and gender. Specifically in Fig. 4(a), the language model perplexity (type II) is shown for language models trained using round-robin for each age and gender group (type II). In Fig. 4(c), inter- and intra-speaker linguistic perplexity is shown for type I models. Error bars depict the inter-

[3] Note that in this experiment only the 132 successful sessions ending in "Arrest", i.e., apprehending the suspect, are analyzed in order to reduce the effect of hesitations due to poor game playing.

[4] Note that, although not shown here, the vocabulary trends and statistics where very similar if stemmed words were used instead of word forms.

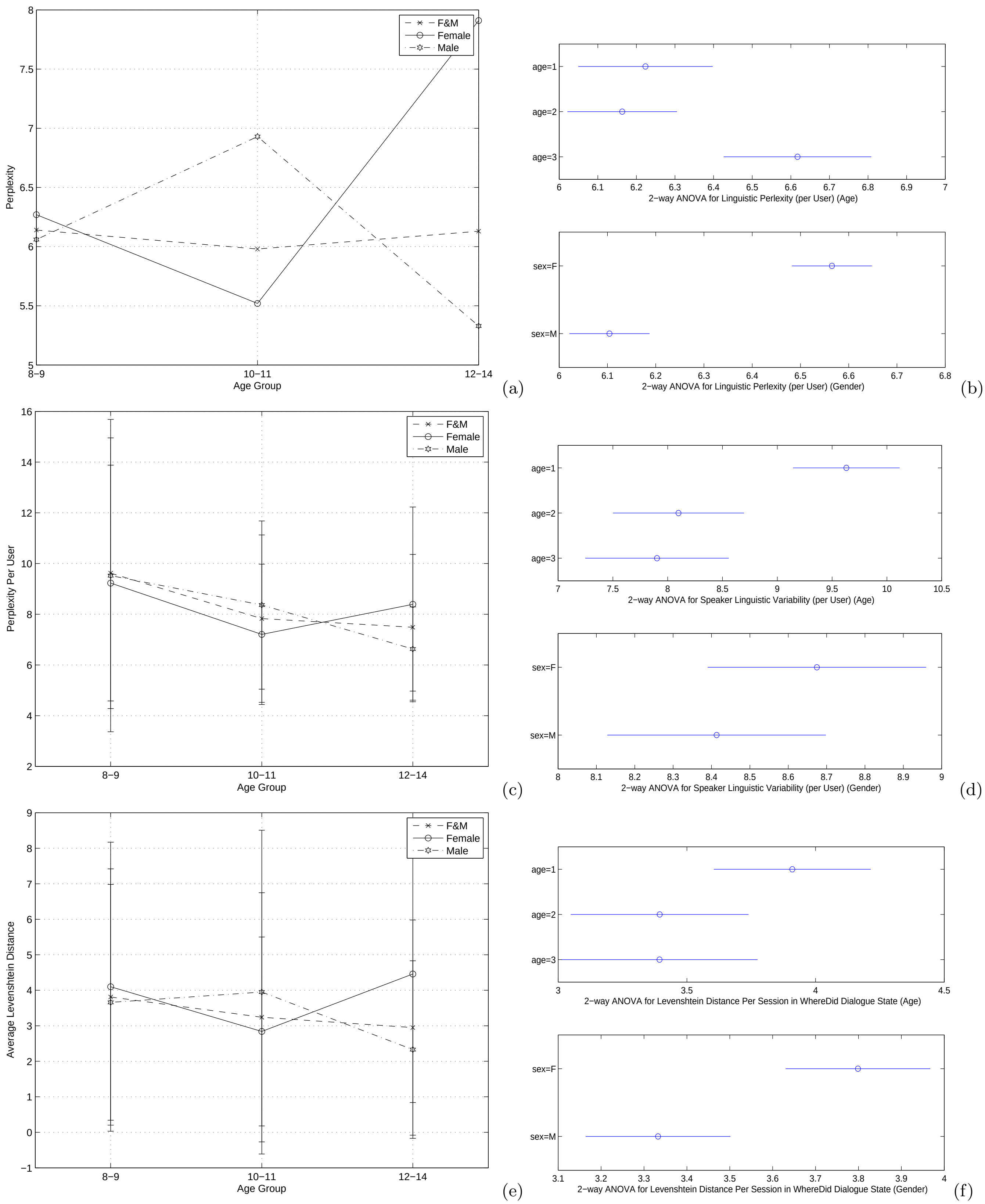

Figure 4: (a) Language model perplexity (type II) per user as a function of age and gender, and (b) 2-way ANOVA results (age-group and gender). (c) Inter- (error-bars) and intra-speaker (plotted curve) language model perplexity (type I) as a function of age and gender, and (d) 2-way ANOVA results (age-group and gender). (e) Average Levenshtein distance between two adjacent utterances of the same speaker from the "WhereDid" dialogue state as a function of age and gender, and (f) 2-way ANOVA results (age-group and gender).

(within groups) speaker variability while the plotted curve depicts the intra- (between groups) speaker variability. Finally, in Fig. 4(e), the average Levenshtein distance between two adjacent utterances of the same speaker from the "WhereDid" dialogue state is shown as a function of age and gender. All three linguistic complexity and variability measures follow very similar trends, i.e., reduction of perplexity as a function of age for boys, reduction of perplexity between the 8-9 and 10-11 groups for girls and then a significant increase for the 12-14 age group. The reduction in perplexity between the 8-9 and the 10-11 age groups is larger for girls than for boys, although, this result is not always statistical significant. Note also that the increase in perplexity between the girls aged 10-11 and 12-14 holds both within- and across-speakers.

6. DISCUSSION

To better interpret the results of this study it is important to also take into account the effect that the task has on the linguistic choices of the children, i.e., the game itself. In general, it is hard to select a task that is both engaging and challenging for a wide range of age groups. In our case, the selected game is rated for children 8 years and older. Results from the exit interview (and the rate of successful games played) indicate that the game is challenging for the 8-9 age group, a good fit for the 10-11 age group, while the game does not provide much of a challenge for the 12-14 age group.

As far as duration and speaking rate metrics are concerned there is a significant difference between the results for read speech reported in [17] and spontaneous speech reported here. Specifically, vowel durations are significantly lower and speaking rate is higher for spontaneous than for read speech. As far as age trends are concerned, the results for read and spontaneous speech are very similar: a 10% relative reduction in average vowel duration can be noticed between the 8-9 and 10-11 age group and then durations level off or increase somewhat. All in all, the children seem to reach adult-level skill at articulation speed around the age of 11 year, and girls seem to be somewhat more adept than boys in the 8-9 age group. Note that in [17] it is claimed that adult-level skills are attained approximately 2 year later (around 13-14 years vs 11-12 in this study), e.g., see speaking rate. This apparent difference between read and spontaneous speech, might have to do also with the maturation of the reading skills of the children, i.e., it could be that the durations observed in read speech are biased by the reading speed of the child. In fact, the higher absolute duration values for read speech could be explained by the additional cognitive load that reading incurs. More experiments in larger corpora are needed to verify these claims.

The absolute values of mispronounciations and filled pauses are higher in human-machine than in human-human interaction, as reported in the literature [4]. 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. Note that the 10-11 age group is fully engaged by the game and find the game most fun. Further research is needed to interpret this result.

In general, the ability of the children to use language efficiently to achieve a task improves with age for all three age groups. 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. Specifically, note that linguistic variability is reduced with age. Also older children keep repeating linguistic patterns that have been successful at achieving the task at hand (see Levenshtein distance). 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, we conclude that *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 game is "easy" for older children, i.e., for girls aged 12-14 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.

7. CONCLUSIONS

The analysis of acoustic, lexical and linguistic characteristics of spontaneous children's speech has shown significant age and gender trends. Average vowel duration was shown to be significantly lower for spontaneous speech compared with 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 read speech. Disfluencies decreased with age and leveled off for the 12-14 age group, while hesitations increased with age and where especially pronounced for girls in the 10-11 age group. Older children used simpler linguistic constructs and shorter utterances to complete the task. An important finding was that girls showed significantly more linguistic exploration than boys, as was evident, by the increase in average sentence length and linguistic perplexity for the 12-14 age group. These findings could impact the design of spoken dialogue systems (and especially interactive multimodal games) for children both at the modeling and at the interface level. For example, it is clear from this studies that girls, 12 and older, consider language as part of the game rather than just a tool to complete the game (as boys do).

Future work will include factor analysis for the acoustic and linguistic measures. The following additional factors will be investigated: exit interview scores, e.g., user satisfaction with the game and emotional state of child, e.g., child frustration. In addition, a simple semantic and pragmatic analysis of the children utterances and statistics will be computed, e.g., number of semantic attributes filled per utterance. Further research in the field of analysis of spontaneous children speech will contribute in both improving the acoustic and linguistic models of children speech and in implementing better interactive dialogue systems for children.

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

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

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

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

[10] 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 the International Conference on Acoustics, Speech, and Signal Processing*, (Toulouse, France), May 2006.

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

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.