SPOKEN DIALOG SYSTEMS FOR CHILDREN

Alexandros Potamianos and Shrikanth Narayanan

AT&T Labs–Research, 180 Park Ave, P.O. Box 971, Florham Park, NJ 07932-0971, U.S.A.
email: {potam,shri}@research.att.com

ABSTRACT

In this paper, we outline the main issues when designing interactive multimedia systems for children and propose a unified approach -acoustic, linguistic, and dialog modeling- to system development. Acoustic, linguistic and dialog data collected in a Wizard of Oz experiment from 160 children ages 8-14 playing an interactive computer game are analyzed and children-specific modeling issues are presented. Age-dependent and modality-dependent dialog flow patterns are identified. Furthermore, extraneous speech patterns, linguistic variability and disfluencies are investigated in spontaneous children's speech, and important new results are reported. Finally, baseline automatic speech recognition (ASR) results are presented for various tasks using simple acoustic and language models.

1. INTRODUCTION

Children and adults differ in how they speak to and interact with machines. From the perspective of interactive spoken dialog systems, such differences can be identified at various system levels, e.g., acoustic and linguistic modeling of speech, dialog strategies and preferred interaction modality. Specifically, the acoustic correlates of children display increased dynamic range and high variability when compared to adults, which can be mostly attributed to vocal tract growth and motor control development of the articulators. Significant linguistic differences exist between children and adults, mostly in the degree of linguistic variability and disfluencies. Furthermore, problem-solving skills and approaches differ widely with age. Finally, the dynamics of man-machine interaction are not necessarily the same for children and adults.

Investigations on the acoustic characteristics of children speech have shown systematic age-dependent variation in acoustic correlates such as formants, pitch and duration [2, 3]. These results have been exploited in developing speaker normalization and model adaptation algorithms to improve automatic speech recognition for children [5]. Nevertheless, basic questions in the field of ASR acoustic modeling of children's speech still remain unanswered. The linguistic aspects of children's speech have not been adequately modeled, especially for spontaneous speech. In addition, little work exists in analysis and modeling of conversational user interfaces for children and in investigating different modalities of child-machine interaction. Previous published work on interactive voice-controlled systems for children focus mostly on educational applications and have very limited scope in terms of providing a natural dialog interface [7, 4, 6].

In this paper, we attempt to characterize the main differences between children and adult speech from the viewpoint of spoken dialog system design. Data are collected from children 8-14 years of age and adults (for reference) when using voice to control an interactive computer game under a Wizard of oZ (WoZ) scenario. The dialog flow data are analyzed and differences in dialog strategies are identified as a function of age, gender and input modality (voice vs. keyboard). Inter- and intra-speaker linguistic variability is compared across different dialog states and speech disfluencies are analyzed as a function of age and gender. This, to our knowledge, is the first comprehensive attempt at linguistic analysis of spontaneous children's speech. Finally, preliminary ASR experiments are performed using simple acoustic and language models. Important modeling issues and guidelines for building ASR systems that are robust to spontaneous children's speech are identified and a unified –acoustic, language and dialog modeling– approach is proposed for system development.

2. EXPERIMENTAL SETUP

The Wizard of oZ (WoZ) experimental setup is shown in Fig. 1. The player sits in front of a slave monitor wearing headphones, i.e., watching and listening to the audio-visual output of the wizard's computer. In the observation room, the wizard controls the experiment interpreting the voice input from the player and taking appropriate action. A separate loudspeaker next to the slave monitor is used to play pre-recorded error-control and clarification messages. High-quality audio recordings of the player's voice commands are collected using a close-talking head-mounted microphone and a far-field microphone (the game audio output is also recorded for reference). A video recording of the "picture-in-picture" image of the player and the game screen complete with the (mixed) audio from player and computer is also obtained.

2.1. Game Description

The software selected for this WoZ experiment was the popular computer game "Where in the U.S.A. is Carmen Sandiego?" (WITUICS) by Brøderbund. WITUICS is an interactive detective game 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 (in one of fifty U.S. states). The player can talk to characters on the game screen and ask them for clues that can be correlated with information in a geographical database. Information can be obtained from the database either by

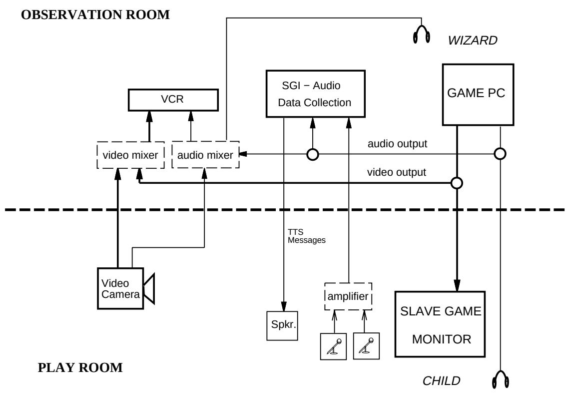

Figure 1: The experimental WoZ setup.

(single or multiple word) search or by stepping through a hierarchical structure of queries. The player needs to create the suspect's profile (five features, each with two to five pre-defined attributes) and issue a warrant when all fields are filled. The player has to travel through five U.S. states tracking the suspect, identify him (using the profile information) from among the cartoon characters in the screen and arrest him.

Overall, the game is rich in dialog subtasks including: navigation and multiple queries, database entry, and database search. Further, the dialog between the player and the machine is natural and human-like because during a substantial part of the game the player converses with cartoon characters on the screen. As a result spontaneous speech can be elicited. Finally the game offers a good opportunity to investigate extraneous speech patterns in children speech.

The structure of the game was not changed (no adaptation to voice modality). The only modifications to the game were providing some degree of automation (by the wizard) when navigating in the top level and scrolling in the database, and adding four text-to-speech generated dialog error control/clarification messages.

2.2. Experiments and Population Statistics

A variety of experiments have been run using voice(V), keyboard and mouse(K+M), or voice, keyboard and mouse (V+K+M) to control the game. A comparison of different modalities for child-machine interaction is beyond the scope of this paper. Overall the children had an overwhelmingly positive impression of using voice to control WITUICS. We will focus on the analysis of the dialog, linguistic and acoustic data when using voice or K+M to control the game. Further we will only report results on interactions with "perfect" speech recognition and "perfect" mapping from recognized speech to game responses/actions.

Prior to each experiment, the game and the voice interface were explained to the player by a moderator (children players were not informed of the existence of a wizard). Most players played two games (23% played a single game and 3% played three games). Data from a total of 160 children and 7 adult players were collected. The total number of games played (using voice with no recognition errors) per age group and gender are shown in Table 1. A total of about 50000 utterances were collected. After the completion of each experiment, the moderator interviewed the subject to gauge the user's perception regarding the game and the interface.

	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

Table 1: Number of games per player's age and gender.

3. DIALOG DATA ANALYSIS

In this section, the results from the analysis of "dialog data" collected when using voice or K+M to control the game are presented. Speech utterances are assigned to dialog states according to the game actions they trigger, and thus "dialog flow" and "sequence of game actions" are equivalent terms (provided that the wizard does not err). Transition between dialog states were analyzed for various game subtasks and age-dependent trends were identified. Next, extraneous speech patterns were identified and their occurrence in dialog flow was modeled. Finally, game flow differences when using voice or K+M modalities were identified. Our goal here is to provide guidelines for designing dialog systems and for choosing the appropriate modality for each dialog subtask.

Dialog states were defined to roughly correspond 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" incorporates 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. Speech utterances were assigned to dialog states while the game was played on the collection machine by a labeler (assistant to the wizard).

As mentioned above, the game dialog flow mainly consists of navigation/query, database search and database entry subdialogs. In Fig. 2 we show the dialog flow diagram for the navigation/query subdialog. The total number of times a state is visited (in parenthesis) and the total number of state transitions (arrow labels) are shown for all games played by children players (total of 290 games). The graphs provide us with useful information about dialog and problem-solving strategies of children. For example, it can be deduced from Fig. 2 that only 30% of the time a game character is asked to provide information for both the location and the physical characteristics of a suspect, i.e., most children prefer to concentrate on a single task. Similar remarks can be made for the frequency of skipping states (e.g., executing a "Find" without opening "Database" in the database search subdialog), or the frequency of superfluous greetings (e.g., "Goodbye").

Gender and age trends in the dialog flow were also studied. Two age groups were used for this purpose: 8-10 and 11-14 year-olds. The dialog flow patterns were practically identical between male and female children. The dialog structure differences between the 8-10 and 11-14 age groups were significant and can be mostly attributed to improved game-playing skills for the older children. On average the children in the 11-14 age group play more efficiently: They complete the game faster, spend less time in database search (more knowledgeable), use more advanced dialog patterns (jump dialog states, merge two commands in one) and attempt multiple goals in a given sub-interaction more often (double queries). Finally, the number of utterances classified as "extraneous speech" were about half as many for the 11-14 age group than for the 8-10 age group.

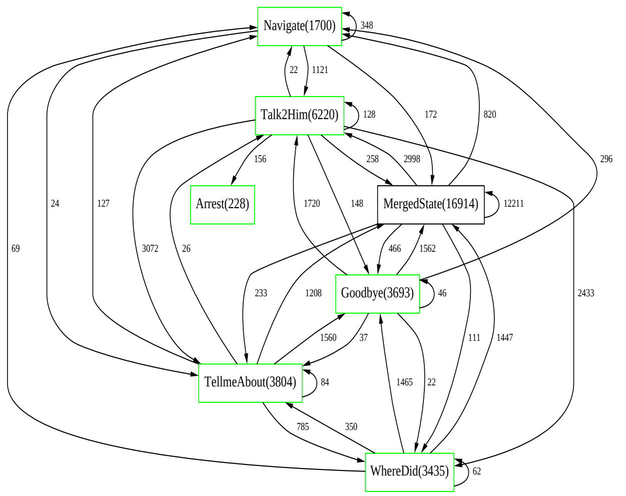

Figure 2: Dialog state and state transition diagram (with counts) for all children players for the navigation/query subdialog ("MergedState" denotes all dialog states not shown in plot).

3.1. Extraneous-speech modeling

We define extraneous speech utterance as any utterance that triggers no valid game response or action. Statistical modeling of (sequences of) dialog states that precede extraneous speech is important for designing robust dialog systems for children. In our data, extraneous speech utterances correspond to approximately 5% of all utterances spoken for the 8-10 year-olds (3.7% for all subjects), varying from 0% to 25% among subjects (7% variance). Most extraneous speech utterances fall in one of the following categories: (i) expressing excitement (disappointment) when vital (useless) information is provided by the game or success (failure) is achieved in one of the game stages, (ii) requesting game-strategy information, interpretation of game output or approval by other people in the room (an adult moderator or other children were present in the game room for about half of games played), and (iii) interacting with characters on the screen irrelevant to game goals and objectives. Overall, the extraneous speech utterances were found to be highly speaker-dependent, age-dependent, and to be preceded by a small subset of dialog states. Extraneous-speech modeling at the dialog level can significantly contribute to successful utterance verification strategies.

3.2. Dialog strategies: Keyboard vs. Voice

A total of 12 children players alternated on using voice and K+M to control the game. The dialog/action flow and underlying task solving strategies were very similar for both modalities. The total number of commands was roughly the same for the navigation/query and database entry subtasks. For the database search subtask about 50% more actions were taken when using K+M than when using voice. This suggests that for the database search task voice is not the most efficient modality (with the current interface). A final observation is that when using K+M superfluous greetings at the navigation/query menu (dialog state: "Goodbye") were reduced by a factor of three compared to using voice. This reinforces the belief that *although voice might not be the most efficient modality, it is the most natural one.*

	Dialog State				
Variability	1	2	3	4	5
Intra-speaker	0.43	0.26	0.22	0.32	0.40
Inter-speaker	1.05	0.48	0.40	0.54	0.72

Table 2: Inter- and intra-speaker linguistic variability for dialog states "Talk2Him" (1), "WhereDid" (2), "TellmeAbout (3)", "Goodbye" (4), "OpenCluebook" (5).

4. LINGUISTIC AND ACOUSTIC ANALYSIS

In this section, we investigate inter- and intra-speaker linguistic variability for groups of utterances that are semantically equivalent, i.e., trigger the same game action. Specifically, the average Levenshtein string distance was computed among all strings belonging to the same dialog and speaker class, and compared with average string distance among all speakers. In addition, the frequency of occurrence of disfluencies and filled pauses were measured for each age group. Finally, average word length of utterances, average utterance duration and speaking rate were measured. The analysis was performed on a subset of the data containing 22422 utterances from 79 children speakers.

Linguistic variability for semantically equivalent sentences was measured for "simple" dialog states (corresponding to a single game action) in the navigate/query and database entry subtasks. All sentences collected from speaker n that belong to the "simple" dialog state k are the elements of class $C_{k,n}$. The intra-speaker linguistic variability[1] for dialog state k is then defined as $(1/\sum_n L_{k,n}) \sum_n \sum_{i,j} d(S_i, S_j)$, where d is the Levenshtein word-string distance (with 0.75 penalty for word insertion/deletions and 1 for substitutions), $S_i, S_j \in C_{k,n}$, and $L_{k,n}$ is the total number of words in $C_{k,n} \times C_{k,n}$. Similarly, inter-speaker linguistic variability is defined as $(1/L_k) \sum_{i,j} d(S_i, S_j)$, where $S_i, S_j \in C_k = \bigcup_n C_{k,n}$. In Table 2, we show the linguistic variability for various dialog states. Overall, *inter-speaker variability is almost twice as high as intra-speaker variability.* This suggests that there is potential gain from building speaker-specific language models or from performing speaker adaptation on the language models. Note also that both the inter- and intra-speaker variability in Table 2 varies a lot among dialog states. Finally, we compare intra-speaker linguistic variability between the 8-10 and 11-14 age groups, and between male and female speakers. Overall, female speakers displayed higher intra-speaker variability by about 10% than male speakers but this trend was dialog state-dependent. Similarly, about 10% increase in linguistic variability was found in the 11-14 age group vs. the 8-10 age group.

Disfluencies and hesitations were analyzed as a function of age and gender. Mispronounciations, false-starts, (excessive) breath noise and filled pauses (e.g., um, uh) were labeled for a subset of the data (22422 utterances). About 2% of the utterances labeled contained false-starts and 2% contained (obvious) mispronounciations. Breathing and filled pauses were found in 4% and 8% of the utterances respectively. No gender dependency was found for any of the disfluencies and hesitations investigated. Further, the frequency of occurrence of false-starts were found to be age-independent. The frequency of occurrence of mispronounciations was almost twice as high for the younger (8-10 yrs.

[1] The formulation of the linguistic variability measure was motivated by the within- and between-cluster scatter matrices of discriminant analysis.

old) age group than for the 11-14 year-olds. Breathing 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. Although disfluencies and hesitation phenomena occur more frequently for children than for adults, from a linguistic and acoustic modeling prospective, they do not present an unsurmountable hurdle for building successful spoken dialog systems for children.

Finally, small differences in duration and average string length were found between the two age groups. No gender or age bias was found in the average utterance length (in words). 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 [3].

5. ASR MODELING AND EXPERIMENTS

In this section, we present some preliminary speech recognition results for various subtasks of the voice controlled WITUICS application. Baseline performance was evaluated for phone and sentence-level ASR tasks. Context independent hidden Markov models (HMMs) using 16 mixtures per state and three states per phone were trained on a subset of the data (6444 utterances collected from 51 speakers). The test set consisted of 2023 utterances collected from 20 speakers. The vocabulary consisted of 761 words.

For the phone recognition task (NO phone grammar) the phone recognition accuracy rate was 46.2%. The relative phone error rate increase for children vs adults is about 25% (adult reference from TIMIT), which is consistent with [5]. The higher error rate for children is mostly due to increased inter- and intra-speaker acoustic variability.

The word accuracy using sentence-level grammars for the whole test set and for three subtasks is shown in Table 3. For this experiment, sentence-level recognition for the entire test data was constrained by a phrase grammar that comprised of the 50 most frequent utterances in the training set. The word accuracy with this grammar is significantly higher for in-vocabulary than for out-of-vocabulary (OOV) utterances: 98.2% vs. 31.3%. Tasks I, II, and III correspond to utterances spoken at the query/navigation, database search, and database entry levels, respectively. The grammars for each of these subtasks were constructed using the top 50 utterances in the corresponding training sets of each subtask. Note that on average about two-thirds of the utterances were OOV under this constrained grammar: 50%, 75%, 90% for tasks I, II and III, respectively. The navigation/query task (Task I), the least complex of the tasks, had the best OOV performance while the database search task (Task III) had the worst. When the size of the grammar was increased by including all utterances from the training set, the number of OOV utterances decreased to 33% (task I), 66% (task II), 82.5% (task III), with an overall word accuracy increase to 78.1%, 59.2%, and 41.4%, respectively. Although very high recognition accuracy can be achieved for in vocabulary utterances a large number of utterances are still OOV especially for subtasks task III and II significantly degrading the overall performance.

Overall, the results are encouraging and provide important directions for language modeling. For example, the query navigation task can be best treated as an action classification problem [1] while the database entry/search tasks can be best modeled by general bigram or trigram models with additional strategies for reducing linguistic variability such as using speaker-dependent and adaptive language models. Better language models together with speaker normalization and model adaptation [5] can provide a spoken dialog system for children with high recognition accuracy.

ASR Task	In Voc	Out Voc	Overall
Sentence (All)	98.2%	31.3%	56.4%
Task I	98.9%	47.9%	73.4%
Task II	94.7%	40.9%	53.2%
Task III	100%	24.7%	32.7%

Table 3: Word accuracy using simple sentence-level grammars for the WITUICS game and various subtasks.

6. CONCLUSIONS

In this paper, dialog, language and acoustic spontaneous speech data collected from children ages 8-14 were analyzed and age-dependent characteristics were identified. Extraneous speech patterns, linguistic variability, disfluencies and problem-solving strategies were among the issues visited in this paper and important new results were reported. Overall, the results of the analysis verify the age-dependent trait of most of the dialog, linguistic and acoustic correlates and reinforces the importance of developing children-specific spoken dialog systems. Further, baseline ASR results for various tasks were reported that show that it is feasible to build successful spoken dialog systems for children using existing ASR technology. In depth analysis of the data can provide further insight into acoustic, linguistic and dialog flow differences between children and adults, and assist in developing successful ASR applications for children.

Acknowledgments: The authors would like to express their sincere appreciation to Dawn Dutton for helpful advice during the early stages of this work; to Jay Wilpon and Ben Stern for their support; to Rick Rose for many useful suggestions; to Matt Einbinder, Christa Lazarus, Roger Barkan, John Baldasare and Pete Ciallela for their invaluable help that made this study possible.

7. REFERENCES

[1] A. L. Gorin, G. Riccardi, and J. H. Wright, "How may I help you?," *to appear in Speech Communication*, 1998.

[2] 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.

[3] S. Lee, A. Potamianos, and S. Narayanan, "Analysis of children's speech: Duration, pitch and formants," in *Proc. EUROSPEECH*, Greece, pp. 473–476, Sept. 1997.

[4] 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.

[5] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Proc. EUROSPEECH*, Greece, pp. 2371–2374, Sept. 1997.

[6] M. Russell et al, "Applications of automatic speech recognition to speech and language development in young children," in *Proc. ICSLP*, Philadelphia, PA, Oct. 1996.

[7] 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.