

Child and Adult Speaker Adaptation during Error Resolution in a Publicly Available Spoken Dialogue System

Linda Bell and Joakim Gustafson

Telia Research, Sweden

`linda.e.bell@telia.se`, `joakim.k.gustafson@telia.se`

Abstract

This paper describes how speakers adapt their language during error resolution when interacting with the animated agent Pixie. A corpus of spontaneous human-computer interaction was collected at the Telecommunication museum in Stockholm, Sweden. Adult and children speakers were compared with respect to user behavior and strategies during error resolution. In this study, 16 adults and 16 children speakers were randomly selected from a corpus from almost 3.000 speakers. This sub-corpus was then analyzed in greater detail. Results indicate that adults and children use partly different strategies when their interactions with Pixie become problematic. Children tend to repeat the same utterance verbatim, altering certain phonetic features. Adults, on the other hand, often modify other aspects of their utterances such as lexicon and syntax. Results from the present study will be useful for constructing future spoken dialogue systems with improved error handling for adults as well as children.

1. Introduction

With few exceptions, spoken dialogue systems developed so far have been designed for adult users. However, a growing number of adolescents and children are likely to access speech based systems in the future. Previous studies have shown that children's voices are more variable in terms of acoustic-prosodic features as well as more disfluent when compared to adult speech [1, 2]. As a consequence, conventional speech recognizers, trained mainly on adult speech, have difficulties handling children's voices [3, 4]. Moreover, it appears that children overall employ partly different strategies when interacting with dialogue systems than adults do. A Wizard-of-Oz study has indicated that younger children use less overt politeness markers and verbalize their frustration more than older children do [5]. It has also been shown that children's user experience is improved if they can communicate with a system with a 'personality' and that they benefit from being able to chose from several input modalities [6]. Recent studies with a simulated system have also shown that children adapt their response latencies [7] as well as their amplitude [8] to that of their conversational partner, in this case different TTS voices.

As speech technology develops, spoken dialogue systems will be used to perform increasingly difficult tasks. The demands on these systems to be robust enough to handle real-life environments and mobile users are also increasing. Most commercial applications developed to date have focused on users in a relatively quiet and controlled environment, such as a home or an office. During the last few years, however, spoken dialogue interfaces intended for public settings have also started to emerge.

Examples are systems for information-retrieval tasks over the telephone and call routing [9-11] and information kiosks in public places [12, 13]. In such systems, the outside environment is a factor which is potentially very difficult for the system to assess and handle.

When children begin to use speech-based systems, it is unlikely that they will want to sit in a quiet office and dictate. Instead, kids will probably want to play portable games or access information in a variety of environments. The challenge involved in developing spoken dialogue systems for children is therefore dual; Children's spoken language is inherently more difficult to handle and the systems used by children will be exposed to real-life settings.

In this paper, we describe a speech corpus collected with the Pixie system. This system invites members of the general public, who are visitors at the Telecom museum of all ages, to engage in spoken dialogue with the animated agent Pixie. So far, the Pixie corpus consists of 35.000 utterances of spontaneous computer-directed speech, out of which 25.000 have been orthographically transcribed and labeled. In this empirical study, we focus on how users of the Pixie system alter certain features of their speech when they encounter a communicative failure. In the previously developed August spoken dialogue system, the general public was also asked to interact with an animated agent. August was an experimental and fully functional system, which was used to collect a database of more than 10.000 spontaneous utterances in Swedish. Studies of the August corpus indicated that adults and children use partly different strategies when their interaction with an animated agent becomes problematic [13, 14]. In this article, differences between children and adult speakers' reactions to errors in publicly exhibited systems are once again examined. When compared to the August system, Pixie has a more controlled setting and well-defined speaker database. This makes it possible to study longer sequences of child and adult dialogue behavior in this fully functional system.

2. The Pixie system

The Pixie system is placed in the permanent exhibition 'Tänk Om' ('What If'), where visitors can experience a full-size apartment of the year 2010. The animated agent Pixie (see Figure 1) with whom the users interact in spoken Swedish is supposed to visualize an embodied speech interface to both information services and home control in this apartment. Visitors enter the exhibition in groups of up to 25. Before entering, they must provide the system with some personal background information such as age and gender. This information is stored in the system's database and simultaneously



Figure 1 Children interacting with the Pixie system at the Telecom museum in Stockholm

encoded into a smart card. To begin with, the visitors watch a film which introduces the apartment and Pixie. Next, they enter the apartment where twelve computer screens have been built into walls and tables, enabling the visitors to interact with Pixie by talking into handheld microphones. With all twelve screens potentially being used at the same time, and people simultaneously speaking to each other, the acoustic environment is very challenging. The visitors are asked to either help Pixie perform certain tasks in the apartment or encouraged to ask the agent general questions about herself or the exhibition. The Pixie agent and a few young users interacting with the system can be seen in the figure.

3. Corpus, coding and analysis

So far, after the first 6 months of recordings, about 35.000 utterances of spontaneous computer-directed Swedish have been collected. Out of these, 25.000 utterances were manually transcribed at the word level as well as labeled with tags for exaggerated pronunciation in terms of loudness and hyperarticulation. In the transcribed database as a whole the number of speakers is 2885, and the average number of utterances per speaker is about nine. The utterances where someone other than the smart card holder appeared to be speaking were tagged with “wrong speaker”. These utterances were excluded from the database.

To be able to examine whether the users’ speaking rate increased or decreased during error handling, all utterances were acoustically analyzed. From the corpus of transcribed and labeled data, we took out 15.000 utterances from dialogues with more than five turns. The segmentation of the speech material into words and phonemes was achieved by means of an automatic alignment algorithm [15]. The input to the auto aligner is a speech file and a verbatim transcription of the speech. The output consists of two tiers marking words in standard orthography, and phonemes, respectively. The phoneme tier is supplemented with lexical prosodic features such as primary and secondary stress and word accent type (i.e. accent I or II). The grapheme-to-phoneme conversion, as well as the lexical prosodic markup was accomplished with the KTH text-to-speech system.

Our main interest in the current study was to examine users’ error handling strategies in longer sequences of human-computer dialogues. To be able to examine other phonetic features and

dialogue strategies in greater detail, we made yet another selection. From the 15.000 aligned utterances, we randomly extracted 16 adult and 16 children speakers whose interactions with the system consisted of between 15 and 25 user utterances. The children in this subsection of the corpus were between the ages of 9 and 12. Both the adult and child groups were gender balanced.

As mentioned above, users engaged in two types of dialogues with Pixie. In the system-driven *domain* dialogues, speakers were asked to help Pixie perform certain tasks in the apartment. In the user-driven *social* dialogues, speakers could ask the agent questions about herself, the home of the future, or the exhibition. The corpus of 32 speakers was also transcribed at the dialogue level with the following tags for each user utterance: *Normal*, *meta*, *error*, *repetition* and *rephrase*. *Normal* were all utterances that were part of the typical interaction with Pixie, *meta* were comments about the system or dialogue itself, *non-cooperative* were utterances in which speakers refused to answer the system questions, *repetition* were verbatim repetitions of the previous utterance and *rephrase* all non-verbatim repetitions and rephrasings. We also labeled the dialogue for how the system’s previous turn had affected user behavior, inserting the following tags on each user utterance: *Correct*, *rejected* and *wrong*. *Correct* was the label used when the user’s previous turn was correctly handled by the system, *rejected* was used when the speech recognition confidence score was too low which led to system prompts such as “I didn’t hear/understand you” and *wrong* was used in the cases where there was a misrecognition which led to the wrong response. In Tables 1a-c below, three examples of labeled user utterances are shown. In these examples, a child and two adults interact with Pixie in the social phase of the dialogue.

Where did the family go?	normal
Where did the family go?	repeat
Do you know where the family went?	rephrase
Do you know where the family went?	repeat
Do you know where the family went?	repeat
Where did the family go?	rephrase

Table 1a A child speaker interacting with Pixie

I want water in the bathtub	normal
Turn on the faucets	rephrase
Turn on the water	rephrase
Turn on turn on the water faucet	rephrase

Table 1b An adult speaker interacting with Pixie

How old are you?	normal
Eh you have big ears where do they come from	normal
Eh well now I don't know what to say but	meta
What do you do in your spare time?	normal
When does the exhibit close?	normal
What sorts of things do you eat?	normal

Table 1c An adult speaker interacting with Pixie

Finally, a subjective measurement of perceived speaking loudness was individually assigned to each utterance. Here, we used the labels *low*, *normal*, *high*, *very high* and *scream*. Subjective measures of *hypo-* and *hyperarticulation* were added, as well as labels for *mispronunciation* and 'silly voices'.

4. Results

Our analyses indicate that adults and children use partly different strategies during error handling with the Pixie system. When the dialogue fails, speakers often make one or several attempts to resolve the problem and make themselves understood. As can be seen in Figure 2, children often repeat the same utterance several times. Adults, on the other hand, tend to rephrase their original utterance instead of repeating it verbatim. In the user-driven social dialogues, this pattern is especially clear. When Pixie had failed to interpret their original utterance correctly, adults would attempt to rephrase it or simply move on to the next query. The latter strategy is exemplified in Table 1c above.

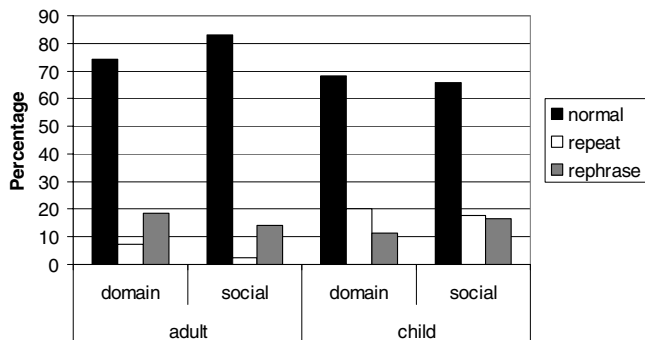


Figure 2 Percentage of all utterances in each category labeled as *normal*, *repeat* and *rephrase*

A closer examination of the utterances labeled as *rephrase* reveal interesting differences between adults and children within this group. When children rephrase their previous utterance they typically add or take away a non-content word. That is, they seldom or never modify the phrase structure or lexical content of the utterance. Table 1a is an example of this type of repetitive sequence, where a child goes back and forth in her efforts to convey her message to the system. When adults rephrase a previous utterance, however, different patterns can be seen. Instead of being near-repetitions, these sequences often contain

greater changes in lexicon and/or phrase structure; see the example in Tables 1b. In this sequence, the adult tries several ways of expressing what she wants done, and modifies her lexical choices repeatedly.

During repetitive sequences, users often modify different acoustic-prosodic features of their speech. As shown in Table 2 below, children's utterances in all categories were judged as hyperarticulated to a higher degree than adults'. The utterances labeled as *normal* were hyperarticulated almost twice as often for children when compared with adults. Half of the adults' verbatim repetitions were hyperarticulated, while the corresponding figure for children is almost three-fourths.

	normal	repeat	rephrase
adults	16%	50%	28%
children	29%	74%	41%

Table 2 Percentage of hyperarticulated utterances for the different utterance types

As can be seen in Figure 3, children more often use loudness as a way of distinguishing repetitions and rephrases from original utterances. While the adults' utterances were never labeled with *scream*, children frequently raised their voices and shouted during their interaction with the animated agent.

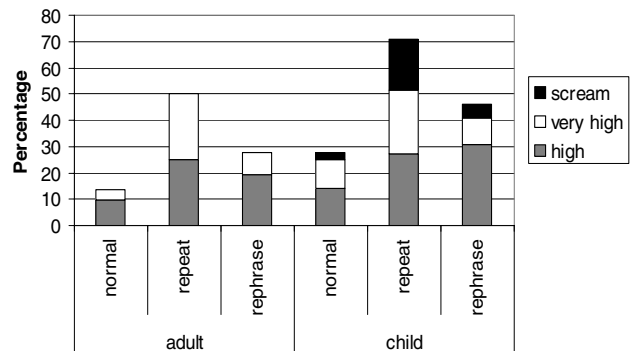


Figure 3 Perceived speaking loudness for different utterance types

In the corpus of 15,000 aligned utterances, verbatim repetitions were compared to original utterances. The automatic alignment algorithm had a high precision at the segmental level. However, the tough acoustic environment with both visitors and other Pixie agents talking in the background made the silence detection quite poor. This meant that both original and repeated utterances had to be manually corrected. All in all, 392 repetitions from children and 387 repetitions from adults were analyzed. The repetitions were on average 29% slower than the original utterances. For children, the repetitions were slightly more than 30% decreased in speaking rate overall. For adults, the repetitions in the domain dialogues were also about 30% slower, while the repetitions in the social dialogues were 22% slower.

The system's previous utterance also affected user strategies in the Pixie corpus. When the system had said "I didn't understand/hear you" in the previous turn, the speaking rate was decreased for both adults and children. Again, this adaptation of speaking rate was more exaggerated for children. This decrease in speaking rate is of course related to the fact that the repeated

utterances often were hyperarticulated. When longer sequences were analyzed, original utterances appearing at the beginning and end of the dialogues were not significantly different in terms of speaking rate. The adaptation of speaking rate that occurs is local, primarily affecting the utterance following a problematic turn.

Children sometimes found it difficult to pronounce some of the options given to them in the system-driven domain dialogues. 11% of the children's utterances in certain difficult dialogue turns contained mispronunciation. Finally, children speakers occasionally modified their manner of speaking by using 'silly voices'. Extreme modifications of voice quality and cartoon-like speech imitations characterize these utterances, which occurred in 4% of all cases.

4. Discussion

Collecting spoken dialogue data in a public environment is a challenging task. It is difficult to maintain control of all variables as large quantities of human-computer dialogues are recorded in a stand-alone system. In the *Pixie* system, the smart cards used for registration and interaction proved to be a robust solution to the problem of assigning user identity, age and gender. In a noisy environment with multiple dialogue systems running in parallel, speech recognition and silence detection becomes problematic. In order to develop a speech recognizer that is robust enough to be used in public, it is necessary to train new acoustic models based on data from real-life settings. However, speech detection and automatic alignment methods are less reliable for this kind of data. Until better models are available, we must use labor-intensive manual methods for preparing speech data from public settings for training purposes.

Several aspects of the *Pixie* system contributed to the patterns of dialogue behavior presently described. The acoustic environment is clearly one such factor, since the general noise level in the exhibition area was quite high. Children visitors often came in larger groups, with many kids simultaneously talking to *Pixie* and other visitors in the room. This can partly explain why children hyperarticulate and raise their voices to such a degree during their interaction with the animated agent. Furthermore, the dialogue design of the system was not tailored for young children and it was sometimes hard for these users to know what to say or (in the system-driven part of the dialogue) to pronounce the options available.

Children and adults react to system errors in different ways. In the repetitive sequences, this can partly be explained by the fact that it appears to be easier for adults to come up with ways of rephrasing their utterances. Children have not yet perfected their language skills, and sometimes fail to come up with an alternative way of expressing a request. It is more difficult for children to modify lexical content and syntactic structure, and they therefore tend to repeat the same thing over and over again or make only minor modifications to their previous utterance. For both adults and children, verbatim repetitions are often hyperarticulated, increased in loudness and longer in duration. However, children's pronunciation is more exaggerated in these respects. Furthermore, when the system fails to understand them adults often move on to the next question in the social dialogues. Children are often persistent in trying to get the system to understand their questions in both domain and social dialogues.

Results from the present study indicate that children and adults use different strategies during error handling with a publicly exhibited spoken dialogue system. Children often repeat the same utterance verbatim several times, while modifying their manner of speaking. Adults also have access to other dialogue strategies, and often adapt their lexicon and/or syntax to meet what they believe to be the limitations of the system. More research is needed to increase our knowledge of the differences between adult and children behavior during error resolution in spoken dialogue systems. For the development of future systems, it is worth considering that children may need more support and guidance when they interact with a spoken dialogue system.

5. Acknowledgements

The work described in this paper was supported by the EU / HLT funded project NICE (IST-2001-35293).

6. References

- [1] Lee, S., A. Potamianos, and S. Narayanan, Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *J. Acoust. Soc. Am.*, 1999. **105**: p. 1455- 1468
- [2] Oviatt, S. and B. Adams, Designing and evaluating conversational interfaces with animated characters, in *Embodied Conversational Agents*, J. Cassell, et al., Editors. 2000, MIT Press: Cambridge, MA. p. 319-343
- [3] Potamianos, A., S. Narayanan, and S. Lee, Automatic speech recognition for children, in *European conference on speech communication and technology*. 1997. p. 2371-2374
- [4] Wilpon, J. and C. Jacobsen, A study of speech recognition for children and the elderly, in *Proceedings of the international conference on acoustics, speech and signal processing*. 1996, IEEE Press. p. 349-352
- [5] Arunachalam, S., et al., Politeness and frustration language in child-machine interactions, in *Proc. Eurospeech*. 2001. p. 2675-2678
- [6] Narayanan, S. and A. Potamianos, Creating conversational interfaces for children. *IEEE Transactions on Speech and Audio Processing*, 2002. **10**(2): p. 65-78
- [7] Darves, C. and S. Oviatt, Adaptation of users' spoken dialogue patterns in a conversational interface, in *Proceedings of ICSLP'02*. 2002: Denver, CO
- [8] Coulston, R., S. Oviatt, and C. Darves, Amplitude convergence in children's conversational speech with animated personas, in *Proceedings of ICSLP'02*. 2002: Denver, CO
- [9] Aust, H., et al., The Philips Automatic Train Timetable Information System. *Speech Communication*, 1995. **17**(3-4): p. 249-262
- [10] Lamel, L.F., et al., The LIMSI RailTel system: Field trial of a telephone service for rail travel information. *Speech Communication*, 1997. **23**(1-2): p. 67-82
- [11] Gorin, A., G. Riccardi, and J. Wright, How May I Help You? *Speech Communication*, 1997. **23**: p. 113-127
- [12] Lamel, L., et al., User Evaluation of the Mask Kiosk, in *Proceedings of ICSLP '98*. 1998. p. 2875-2878
- [13] Gustafson, J. and L. Bell, Speech technology on trial - Experiences from the August system. *Natural Language Engineering*, 2000. **6**(3-4): p. 273-286
- [14] Bell, L. and Gustafson, J. Repetition and its phonetic realizations: investigating a Swedish database of spontaneous computer directed speech, in proceedings of ICPH'99, San Francisco, CA
- [15] Sjölander, K., Automatic alignment of phonetic segments, in *Working Papers 49: Papers from Fonetik 2001*. 2001, Lund University, Dept. of Linguistics: Lund. p. 140-143