Proc. of Intl. Conf. Acoustic, Speech and Signal Proc., Seattle, Washington, May 1998.

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