Creating Conversational Interfaces for Children

Shrikanth Narayanan, *Senior Member, IEEE,* and Alexandros Potamianos, *Member, IEEE*

***Abstract*—Creating conversational interfaces for children is challenging in several respects. These include acoustic modeling for automatic speech recognition (ASR), language and dialog modeling, and multimodal-multimedia user interface design. First, issues in ASR of children speech are introduced by an analysis of developmental changes in the spectral and temporal characteristics of the speech signal using data obtained from 456 children, ages five to 18 years. Acoustic modeling adaptation and vocal tract normalization algorithms that yielded state-of-the-art ASR performance on children speech are described. Second, an experiment designed to better understand how children interact with machines using spoken language is described. Realistic conversational multimedia interaction data were obtained from 160 children who played a voice-activated computer game in a Wizard of Oz (WoZ) scenario. Results of using these data in developing novel language and dialog models as well as in a unified maximum likelihood framework for acoustic decoding in ASR and semantic classification for spoken language understanding are described. Leveraging the lessons learned from the WoZ study and a concurrent user experience evaluation, a multimedia personal agent prototype for children was designed. Details of the architecture and application details are described. Informal evaluation by children was found positive especially for the animated agent and the speech interface.**

***Index Terms*—Automatic speech recognition for children, computer games, conversational agents, human–computer interaction, multimedia interfaces, multimodal systems, spoken dialogue systems.**

I. Introduction

RECENT advances in speech and multimedia technology have spurred worldwide deployment of several prototype and commercial applications that provide natural spoken dialogue with machines [5], [9], [12], [35]. Such efforts are, however, primarily targeted toward the adult user population. While the state of the art in speech technology is still not perfect for the adult population, the task of building spoken dialogue applications for children poses even greater challenges. Children speak and interact with computers differently from adults. Several aspects of these differences can be identified such as in the acoustic and linguistic characteristics of speech, dialog interaction strategies, problem solving skills and user preferences. Further, the sources of these differences reflect physiological and anatomical changes associated with development of articulators and the effects of socio-economic factors during a child's growth. Some of these research challenges will be considered in this paper.

The CHildren's Interactive Multimedia Project (acronym: CHIMP) aimed at providing essential guidelines for engineering successful multimodal-input multimedia-output applications for children with an emphasis on the spoken dialog interface. Factors that motivated this study include 1) children form a crucial segment of customer population for interactive multimedia systems and 2) children are eager and quick to embrace, and use, new technologies. There are several statistics supporting these two facts. More than 60% of children (four to 11 years) use a PC at home compared to 40% for the total U.S. population [17]. Games were found to take up 40% of a preteen's time on the computer, out of a total 6.5–12.5 hours/week, while the rest of the time was devoted to school work [1]. Another study found that out of a total population of about 44 million two to 12 year old children in the U.S., about 25% were projected to be online by the year 2000; the projection increased to 45% for the year 2002 [10]. About 67% of children reported using the internet to gather information, 65% to play games, 49% to do chats, 48% to do creative activities, and 46% to download "stuff" [20]. Similar statistics are available for teenagers. Over 90% of teenagers, across economic boundaries, use computers and more than 70% use the internet [19]. Over 98% of them credit technology for making a positive difference in their lives and 92% believe technology will improve education and job opportunities. Over 71% of them want to *talk* to their computers: speech recognition was the number one high-tech product kids would like to see developed. A third motivating factor for this study was 3) the lack of speech-technology resources for creating voice-enabled applications for children. The resources designed for the adult population are not directly usable for children users. For instance, the age-dependent acoustic and linguistic variability in children's speech [14] makes automatic speech recognition (ASR) for children more difficult compared to adults, hence requiring special algorithms to be designed for providing satisfactory levels of ASR performance. For example, while analyzing ASR performance of the live usage data obtained from their Jupiter spoken dialog system targeted primarily for adult users, Zue *et al.* [35, Fig. 9] found that the in-vocabulary word error rate for children was almost twice that for adult users.

The idea of providing voice interfaces for children's applications is not a new one, however the scope of the systems that have been developed thus far has been relatively limited. Examples of spoken dialog system prototypes for children include word games for pre-schoolers [32], aids for reading [16] and pronunciation tutoring [30]. There is also an increasing number of commercial products being brought to the market—toys

Manuscript received March 28, 2000; revised November 19, 2001. This work was done when the authors were with AT&T Labs-Research, Florham Park, NJ. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hsiao-Wuen Hon.

S. Narayanan is with Integrated Media Systems Center and Department Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: shri@sipi.usc.edu).

A. Potamianos is with Bell Labs-Lucent Technologies, Murray Hill, NJ 07974-0636 USA (e-mail: potam@research.bell-labs.com).

Publisher Item Identifier S 1063-6676(02)01527-4.

and computer games for children—that have limited speech recognition capabilities (small vocabulary, typically isolated or key word recognition). But due to the inherent poor automatic speech recognition and understanding performance, speech has rarely been used as the primary interaction modality in these applications.

Recently, there has also been increasing interest in the design of multimodal interfaces that combine speech with a variety of other input modalities such as text, touch, mouse clicks, handwriting, and gestures [7], [31], [33]. Results of these investigations suggest that the use of multiple modalities, rather than a single modality, leads to more efficient and natural interaction and enhances the overall user experience (for example, [4]). Multimodality is attractive in the creation of conversational interfaces for children in the sense of both overcoming inherent limitations in speech technology and exploiting the ubiquitous availability and/or familiarity with conventional modalities such as the computer mouse, keyboard, joy stick and pen. There are several open research issues that need to be addressed including multimodal input integration and interpretation, multimodal dialog design, multimedia output presentation and performance evaluation. Realistic case studies and prototype designs are crucial to further our understanding of multimodal interactions. The design of a multimodal prototype application for children that will be described in this paper represents an effort in this direction.

Building conversational interfaces for children is a challenging problem and needs to be carried out in several stages. The first step is to establish a proof of concept for the use of speech as a viable means for children to interact with a machine, both in terms of feasibility and usability. Second, data from children need to be collected for quantifying the variability present in their speech and to train and test models for automatic speech recognition (ASR) and spoken language understanding (SLU). This is necessary for ensuring *acceptable* levels of ASR and SLU performance across all ages and environments. Finally, the intuition and results obtained from such data analyses and modeling [24] can be used to create prototype systems.

The rest of the paper is organized as follows. In Section II, acoustic variability issues in children speech and the problem of ASR for children are addressed. The collection, analysis, and modeling of acoustic, spoken language and dialog data obtained from a Wizard of Oz (WoZ) experiment are described in Section III. Results from a user experience evaluation that investigated subjective impressions of children regarding various aspects of the conversational interface are also presented. In Section IV, a design of a conversational multimodal system that leverages the results of the WoZ experiments is presented. The "*Agent CHIMP*" prototype combines speech, keyboard and mouse input modalities and uses text, graphics, speech and animation for output presentation. The application is controlled by animated agents. An informal user evaluation of the prototype and future directions are provided.

II. ASR for Children

Many present day ASR systems, including the ones considered in this paper, use a hidden Markov model (HMM) based pattern recognition wherein statistical models constructed from speech-data samples (training) are used to discern patterns in new unseen speech samples (testing) [28]. Acoustic variability in children's speech, which renders pattern classification difficult, is identified as a major hurdle in building high performance ASR applications for children (Section II-A). In Section II-B the effect of age on the ASR performance of children speech is described. This is followed by a description of how speaker normalization was used to reduce variability and increase the resolution between pattern (phone) classes: A speaker normalization procedure that combines spectral shaping and frequency warping [27] was implemented that resulted in recognition error rate reduction of up to 45%.

A. Acoustic Characteristics of Children's Speech

Investigations of children speech have shown systematic age-dependent variation in the acoustic correlates of speech such as formants, pitch and duration [6], [8], [11], [14]. As a part of the project, changes in the temporal and spectral parameters of children's speech were investigated using speech data (23 454 utterances) obtained from 436 children ages between 5 and 18 years and 56 adults [14]. 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, reaching adult ranges around 13 or 14 years. A specific result that is relevant for ASR is the scaling behavior of formant frequency values with respect to age. As can be seen in Fig. 1(a), the vowel space (boundaries marked by the four-point vowels /AA, IY, UW, AW/ in the F2–F1 plane) changes with increasing age in an almost linear fashion. Also the vowel space becomes more compact with increasing age. 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 with age. Moreover, the first three formants scale similarly especially for males. Females, on the other hand, show a more nonlinear scaling trend for the various formants especially after puberty. The intra-speaker variability was larger for young children, especially for those under 10 years. Fig. 2 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.

There are several implications of these observed age-dependent trends on the ASR of children's speech. The increased spectral and temporal variability in formant values results in greater overlap among phonemic classes for children than for adult speakers, thus rendering the pattern classification problem inherently more difficult. Further, the range of values for most acoustic parameters is much larger for children than for adults. For example, five-year old children have formant values up to 50% higher than male adults [14]. The combination of a large acoustic parameter range and increased acoustic variability seriously degrades ASR performance, as shown in the next section.

Additionally, there are some fundamental issues in processing children's speech. Spectral feature extraction, the typical front-end signal processing step in ASR, is more difficult for children's speech because the fundamental frequency

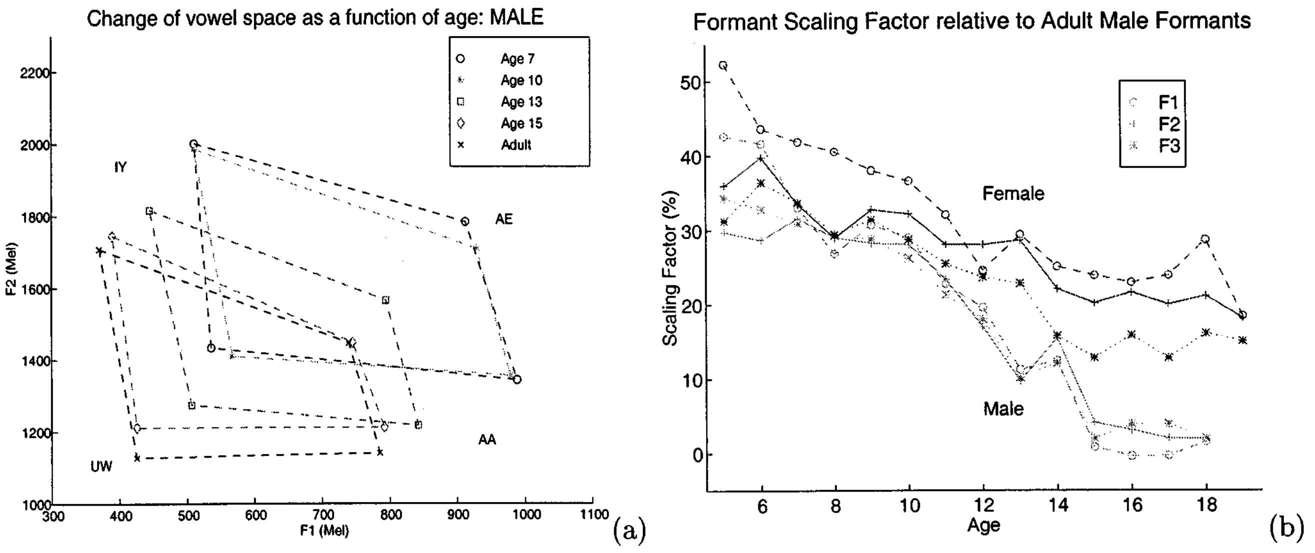


Fig. 1. (a) Changes in F1–F2 vowel space as a function of age. The vowel space boundaries are marked by average formant frequency values for the four point vowels /AA, IY, UW, AE/ for the age groups: seven, ten, 13, 15 and adults. (b) Scaling factor variation in first three formant frequencies with respect to age for male and female children. Scaling was with respect to average values for adult males.

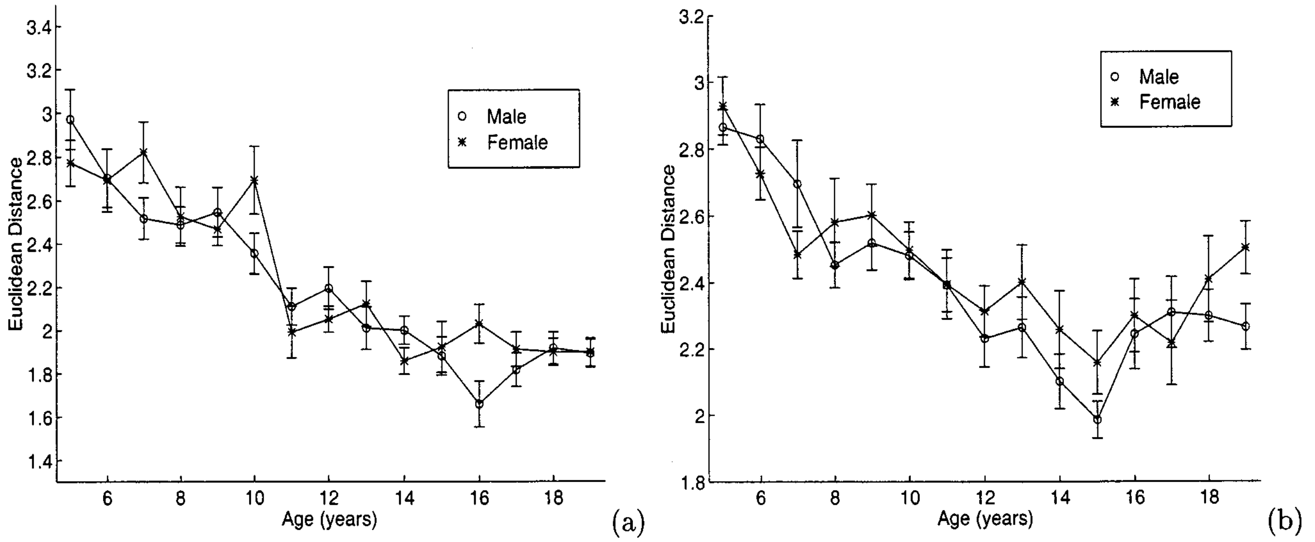


Fig. 2. Intra-speaker, variability as a function of age: (a) mean cepstral distance between the two repetitions of the same vowels and (b) mean cepstral distance between the first- and second-half segments within the same vowel realization.

and the formant bandwidths are of comparable magnitude. Moreover, for a given signal bandwidth (say, a 4-kHz telephony band), there are fewer formants in the spectra of children's speech compared to those of adults. Thus, the sparse sampling of the spectrum (due to high F0 values) and relatively fewer formants in a given bandwidth (due to high formant values) of children's speech pose fundamental limitations on the amount of phoneme-dependent information available at the ASR front-end.

B. Baseline ASR Performance: Children versus Adults

Baseline ASR performance was evaluated as a function of speaker's age for two tasks: 1) connected digit recognition and 2) command and control phrase recognition. The acoustic models for these experiments were trained from speech utterances collected over the public switched telephone network from both adult and children speakers. Details of the training and testing databases are provided in Table I. A mixture of six Gaussians was used to model each state of the context-dependent digit units. Separate phone HMMs for adult and children speakers were trained from the corpora DgtI, DgtII, SubwI and SubwII ("CHLD"), respectively (see Table I for corpus details). A mixture of 16 Gaussians was used to model each state of the 40 context-independent (subword) English phone units.

In Fig. 3(a), word recognition accuracy for a connected digit recognition task (corpus DgtTest) is plotted as a function of age for two model training conditions: models trained from adult speakers (corpus DgtI), labeled "Adult HMM," and from children speakers (corpus DgtII), labeled "Child HMM." For both matched and (especially for) mismatched training and testing conditions, the recognition performance decreases substantially

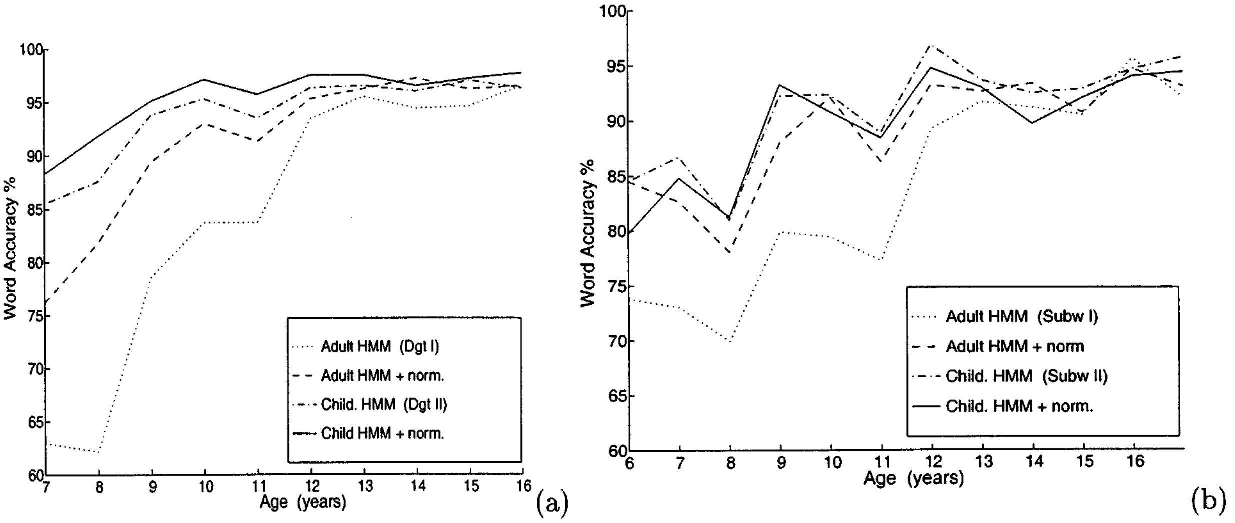


Fig. 3. Word accuracy (%) versus speaker's age using HMMs trained from children ("Child. HMM") or adult ("Adult HMM") speaker population with ("norm.") and without speaker normalization for: (a) connected digit task and (b) command and control task.

TABLE I
TRAINING AND TESTING DATABASES

Name	Speaker Population	Content	No. of speakers	No. of strings
DgtI	Adults	digits	3026	4781
DgtII	10-17 yrs.	digits	1234	5767
SubwI	Adults	phrases	242	12144
SubwII	10-17 yrs.	phrases	1234	14267
DgtTest	6-17 yrs.	digits	501	2656
CommTest	6-17 yrs.	commands	501	3554

for young children. Performance reaches adult levels at approximately 13 or 14 years of age. This agrees with the observation in [14] that by the age of 14 *both* the mean and standard deviation of most acoustic parameters reach adult levels.

Overall recognition performance for children speakers was *up to four times worse* than for adults depending on the speaker's age. For mismatched training and testing conditions ("Adult HMM"), word error rate is approximately two to three times higher than for matched conditions. The major reasons for performance degradation in younger speakers are acoustic mismatch between the training and testing data, increased acoustic variability and the large range of acoustic parameters. Speaker normalization and model adaptation were used to reduce the mismatch and variability. These procedures are summarized in the following section.

C. Linear Frequency Warping and Model Adaptation

The frequency warping approach to speaker normalization aims to compensate for inter-speaker vocal tract length variability by linear warping of the frequency axis by a factor α [13]. Frequency warping is implemented in the mel-frequency filterbank front-end by linear scaling of the spacing and bandwidth of the filters. For each utterance, the optimal warping factor $\hat{\alpha}$ is selected from a discrete ensemble of possible values so that the likelihood of the warped utterance is maximized with respect to a given HMM and a given transcription. Let X^{α} denote the sequence of cepstrum observation vectors warped by a linear frequency warping function. If λ denotes the parameters of the HMM model, then the optimal warping factor is defined as

$$\hat{\alpha} = \arg\max_{\alpha} P(X^{\alpha}|\alpha,\, \lambda,\, H) \tag{1}$$

where H is a decoded string obtained from an initial recognition pass. The selected observation vector sequence $X^{\hat{\alpha}}$ is decoded in a second recognition pass to obtain the recognized string.

There is a large class of maximum likelihood based model adaptation procedures that can be described as parametric transformations of the HMM model or the observation sequence. For these procedures, we let $\lambda_{\gamma} = h_{\gamma}(\lambda)$ denote the model obtained by a parametric linear transformation $h_{\gamma}()$. The optimal parameters of the linear transformation $\hat{\gamma}$ and the frequency warping $\hat{\alpha}$ can be simultaneously estimated. The maximum likelihood criterion can be used to select the appropriate model and also optimize the parameters of the speaker normalization and model adaptation algorithms as follows:

$$\{\hat{\alpha},\, \hat{\gamma}\} = \arg\max_{\{\alpha,\, \gamma\}} P(X^{\alpha}|\alpha,\, \gamma,\, H). \tag{2}$$

The potential of this class of procedures was investigated in the context of speaker adaptation from single utterances. In our case, $h_{\gamma}()$ is a simple linear bias applied to the means of the model distributions or the observation sequence [27], and λ^{n}, $n = 1, \ldots, N$ is a family of age-group dependent acoustic models. The results of speaker normalization and model adaptation applied to the connected digits and command and control recognition tasks are described next.

Experimental Results: In Fig. 3(a), digit recognition accuracies before and after speaker normalization are shown

TABLE II
DIGIT ERROR RATE FOR CHILDREN SPEAKERS BEFORE (BASELINE) AND AFTER SPEAKER NORMALIZATION (NORM.)

Model	Baseline	Norm.	Improv.
Adult HMM	15.9%	8.7%	+45%
Children HMM	6.7%	4.9%	+25%
Cld+Adlt HMM	7.6%	5.6%	+25%

(test corpus DgtTest) for HMMs trained from both adult (DgtI corpus) and children (DgtII corpus) speaker populations. The allowed range of formant frequency scaling was from −20% to +12% and a total of 17 warping factors were examined during frequency warping. The error rate reduction due to speaker normalization was up to 50%, and was greater for young speakers under twelve years of age and when mismatched models trained from adult speakers ("Adult HMM") were used [dotted versus dashed line in Fig. 3(a)]. After speaker normalization, the recognition accuracy for children speakers over 9 years of age was comparable to that of adults. The summary of the cumulative results for all ages is given in Table II. In addition, the performance of an HMM trained from data (equally) mixed from the adult and children corpora DgtI and DgtII is shown (labeled "Cld+Adlt HMM"). Overall, digit error rate reduction by 25–45% was achieved using speaker normalization.

In Fig. 3(b), word recognition accuracy is shown as a function of age for the command and control task are shown for various training and testing conditions. CommI and CommII consist of 10 possible phrases (16 word vocabulary) and 50 phrases (68 word vocabulary), respectively. Similar to the digit recognition task, speaker normalization helped significantly to bridge the gap in performance between the models trained from adult and from children speaker populations. However, recognition accuracy levels obtained for adult speakers were still not obtainable for the younger age group (6–9 years), suggesting that normalization strategies more sophisticated than simple linear frequency warping may be needed.

III. CREATING CONVERSATIONAL SYSTEMS FOR CHILDREN: WOZ EXPERIMENTS

To investigate how children converse with interactive systems and to collect speech data, dialog interaction and user experience data in a realistic spoken language application environment, a Wizard of Oz (WoZ) experiment was designed. Increased acoustic and linguistic variability are typical of spontaneous speech, and the WoZ experiment was aimed at providing valuable data toward evaluating ASR and SLU performance of spontaneous child–machine interactions in realistic scenarios. Note that the ASR performance described in Section II was based on read speech obtained in a relatively controlled set-up.

About 160 children, ages eight to 14 years, participated in the study by playing an interactive computer game using voice commands, or keyboard and mouse control [24]. The software selected for this WoZ experiment was the popular computer game "Where in the U.S.A. is Carmen Sandiego?" (WITUICS) by Brøderbund Software. WITUICS is an interactive detective game for children ages eight years and older. There were several reasons why this computer game was chosen for the study. Overall, the game was rich in dialog subtasks including navigation and multiple queries, database entry, and database search. Further, the fact that (during a substantial part of the game) the child conversed with cartoon characters on the screen made the dialog more natural and human-like. As a result spontaneous speech could be elicited.[1] The structure of the game was not changed (no adaptation to voice inputs). The only modification was the addition of four generic text-to-speech synthesized dialog error control and clarification messages: 1) I can't do that now, what else would you like me to do? 2) Can you spell that for me? 3) What was that? 4) I don't know how to do that. What else would you like me to do?

A. Game Description

To successfully complete the game, i.e., arrest the appropriate suspect, two subtasks had to be completed, namely, 1) determining the physical characteristics of the suspect and completing a profile sketch to enable an arrest warrant, and 2) tracking and apprehending the suspect (by traveling through at least five of the 50 U.S. states every game). The player could talk to various characters appearing on the game screen seeking clues about the suspect's trail and physical appearance. To help interpret the clues thus obtained, the player could use aids such as geographical databases that could be queried using single or multiple word searches. A game was deemed successful when the player traveled to the correct location and identified the suspect correctly (using the constructed profile information) from among several cartoon characters on the screen.

B. Experimental Setup

The Wizard of Oz (WoZ) experimental setup is shown in Fig. 4. The player sat in front of a slave monitor wearing headphones, i.e., watching and listening to the audio-visual output piped from the wizard's computer. In the observation room, the wizard controlled the experiment by providing the appropriate output in response to the user's input. Since the audio-channel of the game was not intercepted, the pre-defined dialog error-control and clarification messages were played through a separate audio channel connected to a loudspeaker placed next to the slave monitor. High-quality audio recordings of the player's voice commands were collected using a close-talking head-mounted microphone (Sennheiser HMD 410) and a far-field desktop microphone (Sennheiser K6-C with a cardioid ME64 capsule). The audio output from the game was also recorded for reference. A video recording of the "picture-in-picture" image of the player and the game screen including the (mixed) audio from player and computer was also obtained. Neither the loudspeaker nor the video camera was reported as being intrusive by any of our subjects.

[1] The children were not informed of the existence of a wizard and an observation room. Further, for approximately half of the experimental runs the player was alone in the game room without a moderator present.

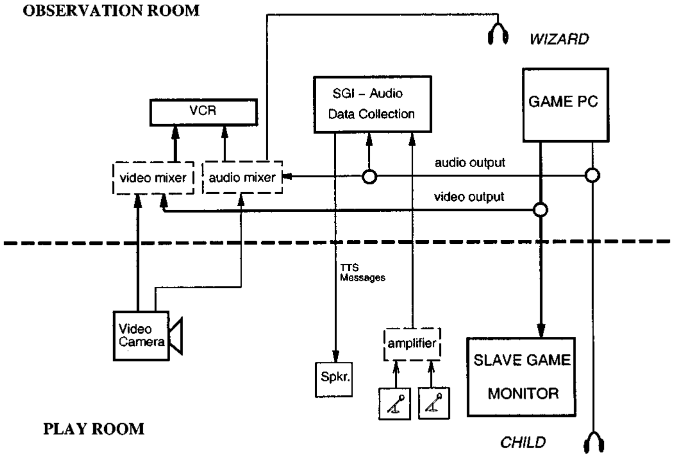


Fig. 4. The Wizard of Oz experimental setup.

C. Experiments and Population Statistics

Although a variety of experiments were conducted using voice (V), keyboard and mouse (K+M), or voice, keyboard and mouse (V+K+M) inputs to control the game, the primary focus in this paper will be on the voice modality interactions.

Prior to each experiment, the game and the voice interface were explained to the child by a moderator. The players were not informed of the existence of a wizard. The wizard followed a set of pre-defined rules during the course of the game while an assistant helped manage the data collection during the experiment. Data from a total of 160 children and seven adults were collected. Most players played two games (23% played one game and 3% played three games). The total number of games played (using voice with no recognition errors) per age group and gender are shown in Table III. A total of approximately 50 000 utterances were collected. After the completion of the experiment, the moderator interviewed the subject to gauge the user's perception regarding the game and the interface.

D. Subjective User Evaluation

All the children who took part in the WoZ experiment participated in an exit interview wherein subjective impressions about the game and the interface were obtained. Sample questions that they were asked include: 1) What did you like about using voice activation? 2) What did you like/dislike about the game? 3) Would you like to use voice input along with keyboard and mouse? The participants were also asked to rate on a scale from 1 to 5 (5 being the highest) the following: voice interface, game, use of headset, TTS-generated error messages, and the use of multimodal inputs.

The children gave very high ratings to the speech interface (93% rated the interface 4 or 5). The game also received high marks but somewhat lower than the interface (only 81% rated the game 4 or 5). The speech interface ratings degraded only slightly when 5% misrecognitions and 5% rejections were randomly introduced into the game by the wizard. It is interesting to compare the relation between the number of games won and the ratings. Losing a game had a significant negative effect on the rating of the game. However, there was no significant effect of the game outcome on the children's rating of the voice input. The 11–12 year olds gave the highest ratings for the voice input. Gender effects were negligible.

TABLE III
NUMBER OF GAMES PER PLAYER'S AGE (IN YEARS) AND GENDER (F—FEMALE, M—MALE)

	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

Other results showed that the dislike for TTS generated error messages and for spelling (for the purpose of "ASR ambiguity resolution") decreased with age. The enjoyment of the use of a headset microphone roughly correlated with the enjoyment of the game. Finally, about two-thirds of the children preferred having a multimodal interface to a voice-only interface. In summary, user experience results were promising for the inclusion of voice as one of the interaction modalities in the design of interactive applications for children.

E. Dialog Data Analysis

In this section, analysis of "dialog" i.e., user–system interaction data, is presented for the voice and keyboard–mouse modalities. Speech utterances were manually assigned to dialog states according to the game actions they triggered [24]. Dialog states were defined to roughly correspond to one (or a group of similar) actions taken by the wizard in response to a user input. For example, the dialog state "Talk2Him" incorporated user queries asking for a cartoon character's attention, while states "WhereDid" and "TellMeAbout" corresponded to queries about the suspect's whereabouts and physical characteristics, respectively. Spoken utterances were assigned to predefined dialog states by the wizard's assistant while the game was being played and were later verified by a group of human labelers. A sample interaction illustrating dialog state tagging is given in Table IV. A total of about thirty dialog states were identified for this application.

The flow through the task was characterized by a sequence of dialog state transitions. The game dialog flow primarily consisted of navigation/query, database search and database entry subdialogs. Fig. 5 shows 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). Such graphs provided useful information about problem-solving and dialog strategies of children. For example, consider the state marked "Tellme-About" in Fig. 5. It can be seen that only about 20% of the time (785/3804) the child requested a second piece of clue; instead, the child preferred to utilize the first piece of information obtained about 72% (2768/3804) of the time this state was visited. In other words, most children preferred to concentrate on a single task per turn. Similar remarks can be made for the frequency of skipping states e.g., executing a database

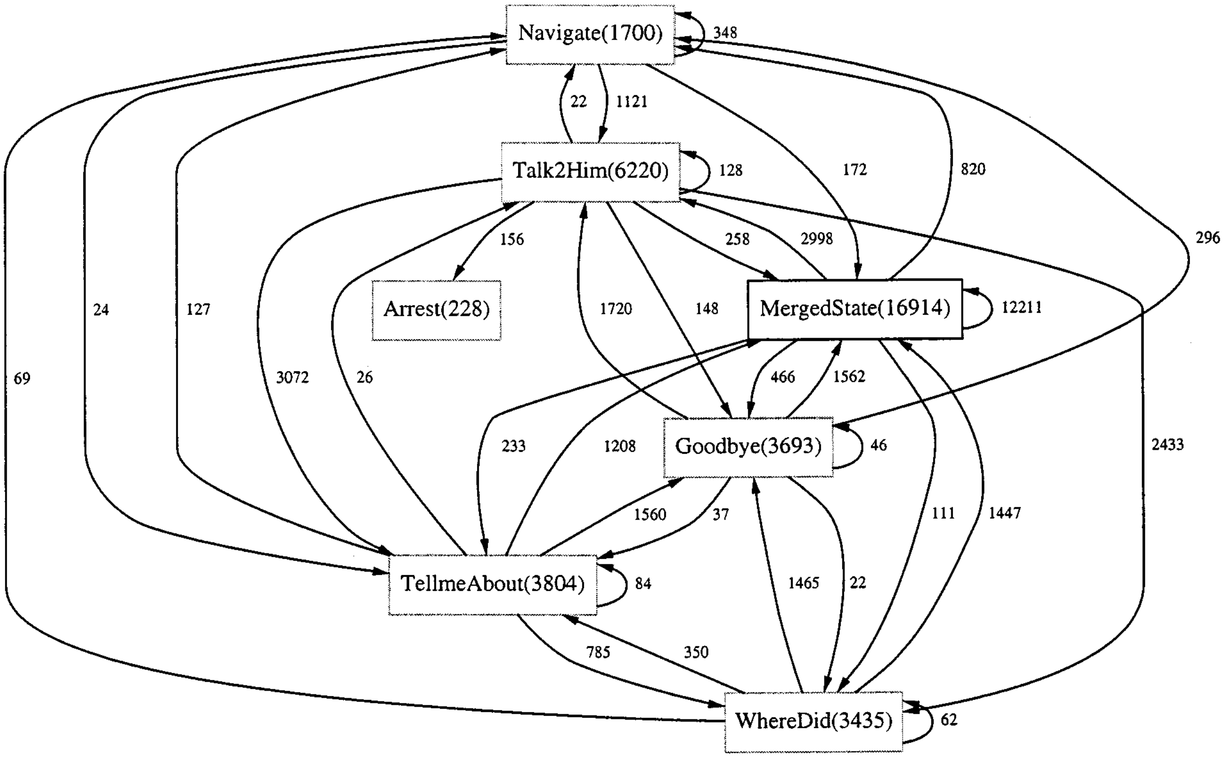


Fig. 5. Dialog state and state transition diagram (with counts) for all children players for the navigation/query subdialog ("MergedState" denotes combination of all dialog states not shown in plot).

TABLE IV
TRANSCRIPT OF A SAMPLE INTERACTION ALONG WITH DIALOG STATE TAGS

User input/System output	Dialogue State
User: Tell me about the suspect? System: She is neither long- nor short-legged	S_{t-3}: TellmeAbout
U: Her height is average S: ... [updating suspect's drawing]	S_{t-2}: EnterFeature
U: Where did the suspect go? S: She is picking peonies in Bloomington	S_{t-1}: WhereDid
U: Go to Indiana S: ... [travel theme]	S_t: GoToState

query (state "Find") without first opening the database (state "Database") in the database search subdialog, or the frequency of superfluous greetings (e.g., "Goodbye"). For example, almost 40% of the transitions out of the in the "TellmeAbout" state (Fig. 5) went through the "Goodbye" state.

Age and speaker dependencies in dialog state transitions were analyzed. Key observations regarding spoken interactions included the following: 1) queries seeking multiple attributes were far less common than those seeking a single attribute 2) frequency of skipping states in the canonical game structure was low 3) frequency of superfluous commands such as "goodbye" was relatively high. There were no noticeable differences in the dialog patterns of male and female children. However, the dialog patterns of older children (11–14 years) were different from those of younger ones (eight to ten years). The older children tended to complete the game faster, did fewer database lookups, used more advanced dialog patterns, and had fewer out-of-domain utterances (about half the number as the younger group).

F. Dialog Strategies: Keyboard and Mouse versus Voice

A total of 12 children players alternated on using voice and keyboard–mouse (K+M) to control the game. In this experiment, each child was assigned to play at least one complete game using voice and one game using keyboard–mouse. The order of which modality was used for the first game was random. The dialog/action flow and underlying task solving strategies were very similar for both voice and K+M modalities. The total number of commands was roughly the same for the navigation/query and database entry subtasks. However, children took fewer turns (almost 50%) using keyboard and mouse than voice to carry out the relatively high-perplexity database search and retrieval tasks. 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 speech might not be the most efficient modality always, it is a more natural modality.*

TABLE V
INTER- AND INTRA-SPEAKER LINGUISTIC VARIABILITY, MEASURED IN TERMS OF NORMALIZED LEVENSHTEIN DISTANCE, FOR DIALOG STATES (1) TALK2HIM, (2) WHEREDID, (3) TELLMEABOUT, (4) GOODBYE, (5) OPENCLUEBOOK

	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

G. Linguistic and Acoustic Analysis

In this section, inter- and intra-speaker linguistic variability for groups of utterances that are semantically equivalent (i.e., those that trigger the same game action) are investigated. Specifically, the average Levenshtein string distance was computed among all strings belonging to the same dialog state and speaker category, 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.

1) Linguistic Variability: Linguistic variability for semantically equivalent sentences was measured for "simple" dialog states (corresponding to a single unambiguous game action) in the navigate/query and database entry subtasks. This subset of the data contained 22 422 utterances. All sentences collected from speaker n that belonged to the "simple" dialog state k were deemed elements of class $C_{k,n}$. The intra-speaker linguistic variability for dialog state k was 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 was defined as $(1/L_k) \sum_{i,j} d(S_i, S_j)$, where $S_i, S_j \in C_k = \bigcup_n C_{k,n}$. Table V shows 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 V varies considerably among dialog states. Finally, intra-speaker linguistic variability was computed for the eight to ten 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, an increase in linguistic variability of about 10% was found in the 11–14 age group versus the eight to ten age group.

2) Extraneous-Speech Modeling: Any speech utterance that triggered no valid game response or action was defined to be extraneous i.e., out of domain. Statistical modeling of (sequences of) dialog states that precede extraneous speech events is important for designing robust dialog systems for children. In the WITUICS data, extraneous speech utterances corresponded to approximately 5% of all utterances spoken for the eight to ten year-olds (3.7% for all subjects), ranging from 0% to 25% among individual subjects (7% variance). Most extraneous speech utterances fell in one of the following categories:

1) those expressing excitement/disappointment when vital/useless information was provided by the game or success/failure was achieved in one of the game stages;
2) those 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);
3) 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. Results imply that modeling extraneous-speech at the dialog level can significantly contribute to successful utterance verification strategies.

Disfluencies and hesitations in the speech data were analyzed as a function of age and gender. Mispronounciations, false-starts, (excessive) breath noise and filled pauses (e.g., um, uh) were manually labeled for a subset of the data (22 422 utterances). About 2% of the labeled utterances contained false-starts and 2% contained (obvious) mispronounciations. Breathing and filled pauses were found in 4% and 8% of the utterances, respectively. While no gender dependency was found for any of the disfluency measures, there was a distinct age dependency. The frequency of mispronounciations was almost twice as high for the younger (eight to ten years) age group than for the older group (11–14 years). 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 11–14 age group. Although disfluencies and hesitation phenomena occur more frequently in children than in adults, our experience showed that ASR performance does not suffer significantly due to these effects, hence requiring no special acoustic modeling strategies.

Finally, small differences in duration and average string length were found between the young and old age groups. No gender or age bias was found in the average utterance length (in number of 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 [14].

H. ASR and SLU Performance—WITUICS Task

Baseline ASR performance and the effects of speaker normalization and model adaptation for connected digit and command/control phrase recognition tasks using read speech from children were described in Section II-B. In this section, ASR and spoken language understanding (SLU) performance for the conversational WITUICS task are presented.

For interactive task-oriented applications such as command and control, unlike dictation applications, it is not always necessary to recognize and understand every spoken word. The SLU problem for the WITUICS task was defined as identifying the next dialog state (ACTION classification) and the values of any associated attributes (ATTRIBUTE recognition), given an utterance transcription. For example, SLU for the utterance “I’d like to go to Indiana” will result in (ACTION: travel, ATTRIBUTE: Indiana) while SLU for the utterance “I’d like to travel” will produce (ACTION: travel, ATTRIBUTE: null). In [26], [29], a unified maximum likelihood probabilistic framework for performing both ASR and SLU was proposed and applied to the WITUICS task. The approach and the results are summarized below. The joint likelihood maximization for acoustic decoding and SLU can be given as

$$\begin{aligned}
&\max_{S_t,\, W_t} P(S_t,\, W_t|O_t,\, S_1..S_{t-1})\\
&\quad= \max_{S_t,\, W_t} P(O_t,\, |W_t,\, S_t,\, S_1..S_{t-1})\\
&\qquad \cdot P(S_t,\, W_t|S_1..S_{t-1})/P(O_t|S_1..S_{t-1}) \qquad (3)\\
&\quad\equiv \max_{S_t,\, W_t} P(O_t|W_t)P(W_t|S_1..S_t)P(S_t|S_1..S_{t-1}) \qquad (4)
\end{aligned}$$

where S_t is the dialog state, W_t is the transcribed user input and O_t is the acoustic observation sequence at dialog turn t. Since S_t is not known at decoding, $P(W_t|S_1..S_t)$ is approximated by $P(W_t|S_1..S_{t-1})$. For computation, the problem is decomposed into acoustic-language decoding and understanding from transcription, i.e., the posterior probability is maximized first with respect to W_t and then with respect to S_t

$$\hat{W}_t = \arg\max_{W_t} \underbrace{P(O_t|W_t)}_{Acoustic}\underbrace{P(W_t|S_1..S_{t-1}}_{Language} \qquad (5)$$

$$\hat{S}_t = \arg\max_{S_t} \underbrace{P(\hat{W}_t|S_1..S_t)}_{Understanding}\underbrace{P(S_t|S_1..S_{t-1})}_{Dialog}. \qquad (6)$$

Both the language model and the dialog model components in (5) and (6) were specified by N-gram automatons. The understanding model (6) in this context is equivalent to determining the next game action i.e., the next dialog state and state attributes. Hence, the understanding model is also specified by dialog class-dependent phrase level N-gram automatons.

To evaluate ASR and SLU performance, the WITUICS data were partitioned into a training set with 6039 utterances (59 speakers, 102 interactions) and a test set with 2050 utterances (21 speakers, 37 interactions). The data included orthographic transcriptions of the spoken utterances and manually-assigned dialog state tags. The acoustic models were context-independent phone models with 16 mixture Gaussians while the language models were dialog state-dependent word trigrams [29]. Bigrams for both the dialog model and the understanding models provided the best SLU classification results. SLU classification accuracy from true transcriptions (obtained manually) was 94.4% while the combined ASR and SLU yielded 86.3% correct classification. Overall, attribute recognition was comparable to the overall ASR word accuracy level of 78%. In summary, 5%–20% error rate reduction came from language adaptation, and 15%–25% from dialog modeling. It was also shown in [26] that further improvements in SLU—of about 10% for this task)—can be obtained by utilizing acoustic confidence scores in the understanding model. The results indicated the feasibility of building viable spoken dialog systems for children using these ASR and SLU technologies.

IV. Building a Prototype

In this section, a case study of designing a conversational multimodal prototype application for children is presented. The design exercise utilized the algorithms, data and results presented in Section II and leveraged the lessons learned from the WoZ study described in Section III. A personal communications assistant (providing telephony, web access and email) and a computer game application for children were used as a vehicle to achieve the following goals:

1) define a general conversational multimodal system architecture;
2) investigate means for merging multimodal inputs (keyboard text, mouse clicks, voice) and multimedia presentation strategies;
3) demonstrate the concept of agent and sub-agent embodiments that handle different modules and functionalities within an application;
4) demonstrate the role of intelligence and personality of the user interface through spontaneous conversation, audio, animation (gestures), and graphics.

While the user interface design focused on children, the system architecture itself was generic. Since the chosen application for prototyping was different from the one in the WoZ study, language data were not readily available. Hence, corpus-driven language and understanding modeling could not be applied at the initial stage of the application creation.

A. System Building Blocks

Fig. 6 shows the main functional building blocks of the system from a user’s perspective. The central part of the system is the controller. The user interface enables interactions using voice, typed text, mouse clicks or combinations there of. Output to the user is presented through audio, graphics, animation and textual modalities. The speech and language processing unit, comprising the ASR and SLU components, enables spoken language interactions. The dialog manager also communicates with information resources such as databases. Further details of the various modules are given in the following sections.

The prototype system consisted of the following components: input/output (I/O) event handler, dialog manager, graphical user interface (GUI), spoken language understanding (SLU), speech recognizer (ASR), speech synthesizer (TTS), animator and database. The speech recognizer used children-specific acoustic models that were built using the data obtained from the WoZ study described in Section III. Since language data were unavailable at the time of the creation of this prototype application domain, application-specific finite state grammars were hand crafted to boot-strap the language models. ASR was performed using the AT&T Watson speech recognition engine.

The dialog manager defines the strategies and actions to be taken based on the user’s input and decides what to present to the user. The dialog manager used in the CHIMP prototype was based on AMICA (AT&T’s Mixed Initiative Conversational Architecture) which provided a library of dialogue actions and a

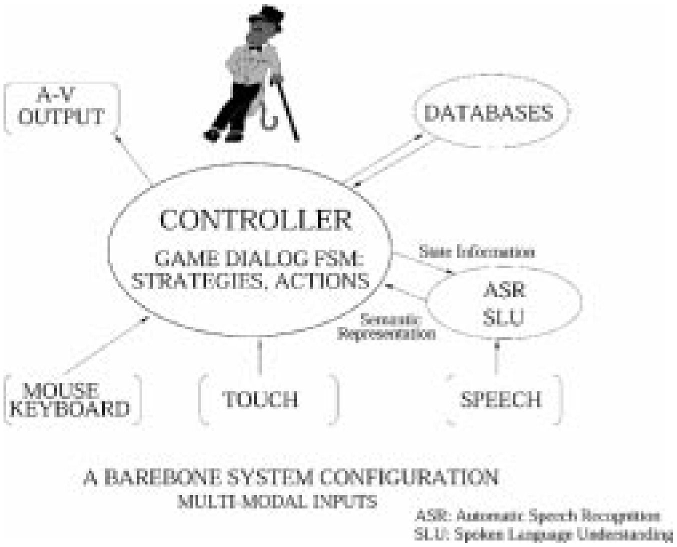

A-V OUTPUT
DATABASES
CONTROLLER
GAME DIALOG FSM: STRATEGIES, ACTIONS
State Information
ASR SLU
Semantic Representation
MOUSE KEYBOARD
TOUCH
SPEECH
A BAREBONE SYSTEM CONFIGURATION
MULTI-MODAL INPUTS
ASR: Automatic Speech Recognition
SLU: Spoken Language Understanding

Fig. 6. Functional block diagram of the CHIMP prototype.

means for specifying dialog strategies either through a JAVA-based GUI or through a high-level scripting language called DMD [22]. The "template" is a data structure that AMICA uses to maintain the dialog state information. The dialog manager can communicate with external modules such as ASR and databases through TCP/IP socket connections.

A semantic representation corresponding to the spoken utterance was derived by performing a lexical analysis (CHRONUS tools, [21]) followed by a rule transduction on the resulting lexical lattice. Due to the lack of in-domain data, rules for deriving semantic representation were boot-strapped by hand-crafted rules in a finite state machine representation. The result of the semantic analysis was represented in a template form. The template structure that provides an "attribute-value" mapping of concepts for use by the dialog manager was derived using a template generator. For simplicity and consistency sake, all actions underlying mouse button clicks and touch on the GUI were also mapped to equivalent, semantically unambiguous, prototype natural language expressions and are henceforth handled by the semantic analyzer in a mode-independent way similar to spoken or typed inputs.

The GUI consisted of five main areas (Fig. 9): graphics/animation area, text area, buttons area, command line area and user command echo area. The GUI design aimed to provide a consistent look and feel across various applications. The personality and appearance of the animated agents provided orientation for the user. Function buttons provided an alternate means of accomplishing several key commands that could also be achieved through voice or typed inputs. A history of user inputs was maintained and any previous input could be easily repeated by highlighting and clicking on the desired entry.

In summary, the input event handler synchronizes the (asynchronous) input from the user (speech, keyboard or mouse events). All inputs are transmitted to the understanding system by way of the dialog manager which, in turn, returns a template with the semantic representation. Based on this semantic representation, the dialog manager decides on the next action to take, resulting in an output template which contains commands for the multimedia presentation. The output template is parsed by the output event handler to yield a multimedia presentation of the system's response (text, graphics, animation, speech). Further details on the communication between the various modules is given in the next section.

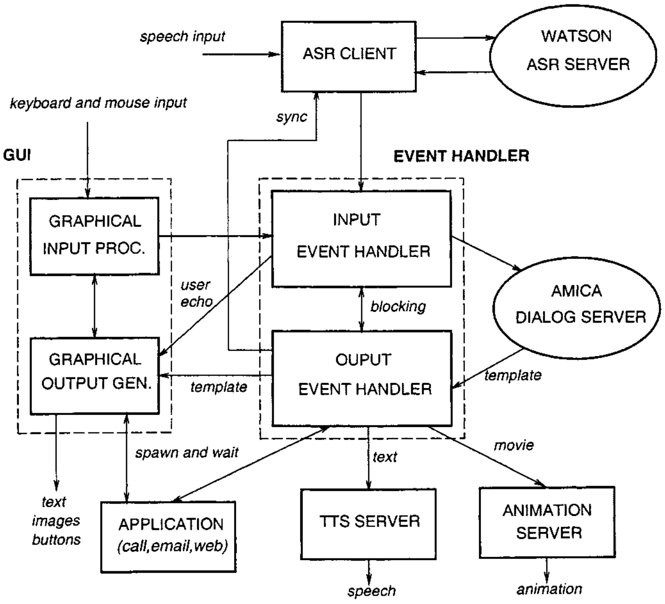


Fig. 7. Architectural diagram of CHIMP prototype.

B. Architecture

An architectural diagram is shown in Fig. 7, where the focus is on the communication between the controller and various servers. The controller is an asynchronous I/O event handler and comprises two separate event loops for the input and output modalities, respectively. The program flow of the controller event loops is straightforward: 1) input events are sent to the dialog manager in the form of text strings (input loop) and 2) templates received from the dialog manager are parsed for multimodal output fields and passed on to the appropriate output modules (text-to-speech synthesizer, animation/movie player or the graphical user interface). During processing of requests by the dialog manager all input events are queued in a stack. If multiple events have been queued up in the input, only the latest event is sent to the dialog manager for processing. The output event handler has the additional functionality of being able to surrender control or simply start up external applications (e.g., pop up an e-mail reader or a web browser). Finally, the controller handles multimodal barge-in events (speaking- or typing-over voice prompts or animation sequences) from the input clients by informing the text-to-speech and movie player modules to stop playback of speech and/or animation sequences.

The dialog manager processes incoming strings by following a control language specification. Strings are interpreted by calling the understanding module and domain information is retrieved from a SQL backend database server. State information and state history are encapsulated in a template form. Based on the current input and the dialog manager specification, the state of the dialog manager gets updated and generates an appropriate response to the user.

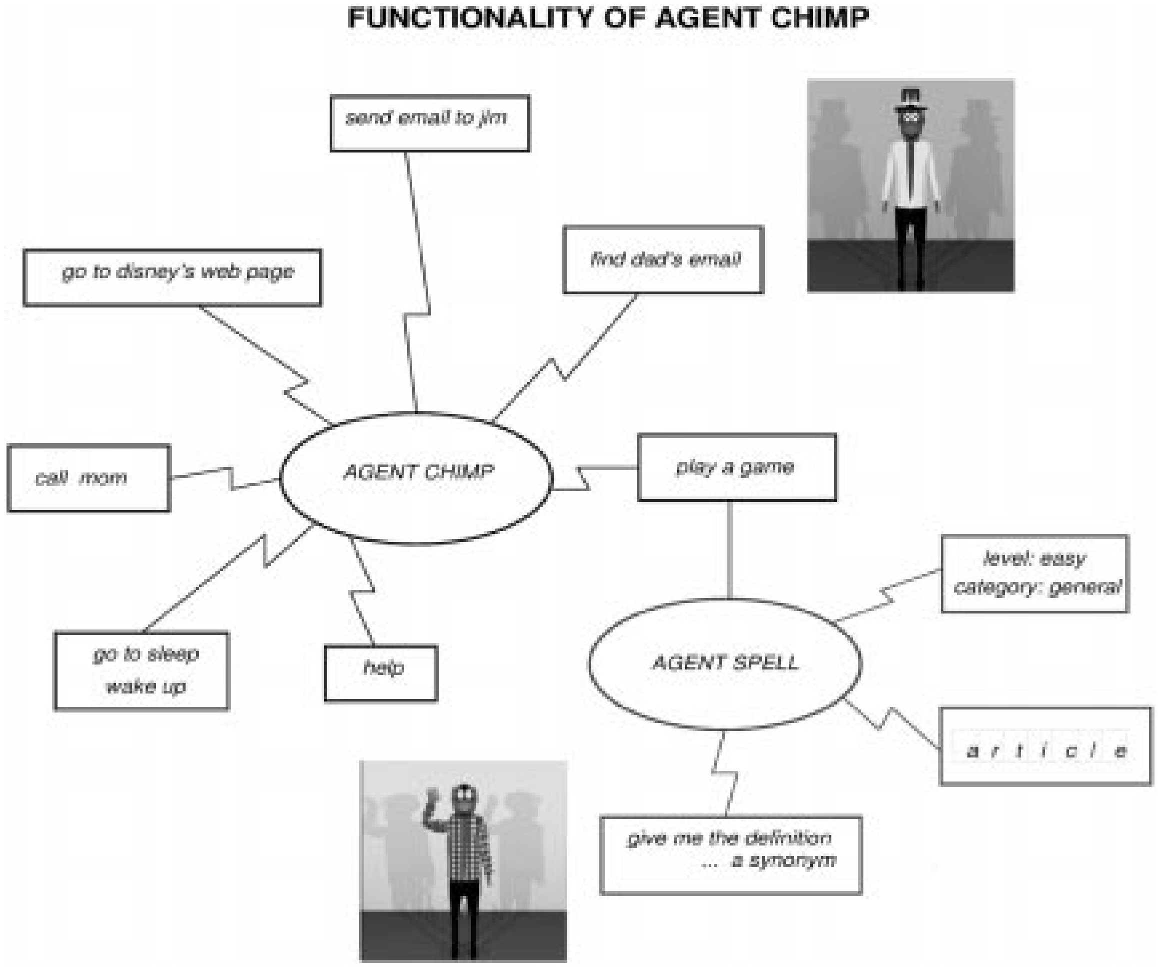


Fig. 8. Overview of some of the application features: Agent Chimp handling the personal agent task and Agent Spell handling the spelling game.

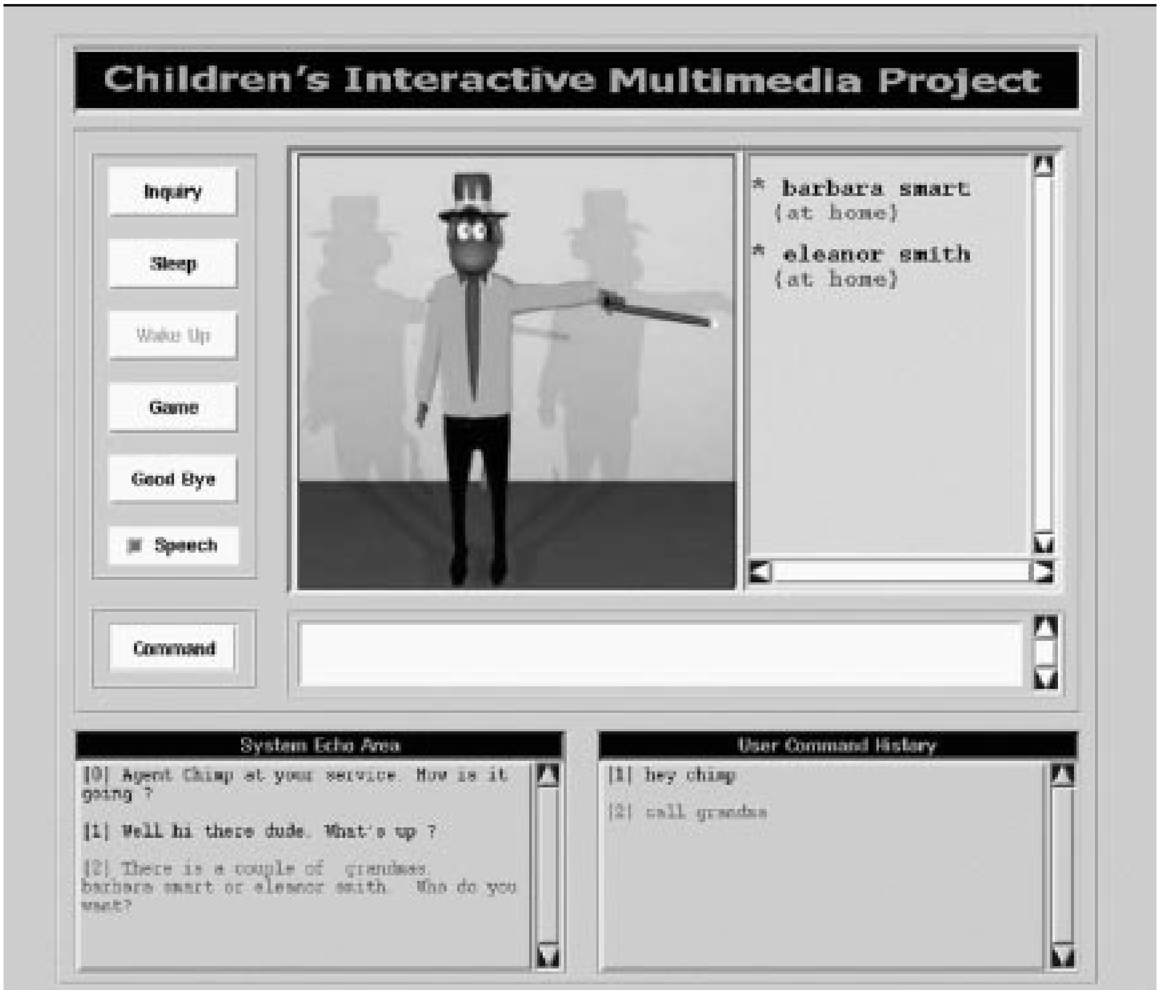
The speech input is processed by the ASR client which communicates with the AT&T ASR WATSON server through a wireline protocol. The commands and audio flow from the client to the server while the results and notifications flow the other way. The interaction between the ASR client and server is through polling or notification modes. Input commands from other modalities are queued in a stack and processed in a similar fashion.

An important principle followed in the design of this prototype is that all the dialog state information is maintained internally in the template that is generated and updated by the dialog manager. In an initial implementation, the state information was transmitted to the I/O handler which in turn decided the form of the multimedia presentation. This strategy was found to be inefficient in the sense that there was duplication of program control logic at the dialog manager and the I/O handler. In a subsequent implementation, the output generation module was folded into the dialog manager: only specific commands for speech synthesis, animation and text display were transmitted to the I/O handler using a predefined protocol. A detailed discussion of the multimodal architecture can be found in [7].

C. Application Details

Tasks: The prototype consisted of two distinct tasks as shown in Fig. 8: a communications agent application (information retrieval from a personal directory, placing phone-calls, accessing the Internet, sending email) and a computer game (spelling bee). Both applications were controlled by a conversational animated agent. The agent embodiment chosen for the CHIMP prototype was a cartoon chimpanzee character. The personality and appearance of the agent, however, were different across the two applications. The user could freely switch back and forth between the applications at any point during the interaction. The system also has "go to sleep" and "wake up" features by means of which the personal agent's attention can be controlled. The personality, including speaking style, and the appearance of the agent for the communications task, Agent Chimp, was designed to be a courteous and sophisticated personal agent. Agent Spell, who handled the spelling game, on the other hand, was a caricature of a studious personality.

The spelling game provided a richer and more challenging application domain than the somewhat simpler command and control nature of the personal assistant task. One of the design objectives was to explore how to design interactive educational tutors for children and provide a test bed for investigating automatic dialog strategy adaptation during the course of an interaction. Agent Spell operated in an "intelligent" mode and provided appropriate feedback and guidance to the child user depending on how the game was progressing. The game progress was conveyed through a score meter on the GUI that kept track



* barbara smart
{at home}

* eleanor smith
{at home}

[0] Agent Chimp at your service. How is it going ?

[1] Well hi there dude. What's up ?

[2] There is a couple of grandmas barbara smart or eleanor smith. Who do you want?

[1] hey chimp

[2] call grandma

Fig. 9. Screen shot of the prototype graphical user interface.

of the number of attempts per word and overall scores with appropriate accompanying audio prompts. The child could take initiative and choose to play a spelling game at the difficulty level of choice (easy/medium/hard) and in a subject area of choice.

Dialog Features: The dialog manager design supported both mixed-initiative (wherein both the agent and the user can take initiative) and system-initiative strategies. Ambiguity resolution (for example, in the case of the retrieval of multiple records) and error control (using confidence measures provided by utterance verification) were implemented as a part of the dialog strategy. Help messages were available through GUI and through audio prompts. Animation provided a key presentation modality particularly in providing both an engaging interface and useful orientation about the current dialog state. The agent personality transformation when the user switched from one task to another (game to communications agent or vice versa) provided task level orientation. This was accomplished by an animation sequence that consisted of the old agent personality disappearing behind a dropping curtain and the new personality emerging when the curtain was opened. The following meta dialog presentation features were included: pointing gesture for information presentation (particularly to indicate the desired selection from among a list), shrugging for conveying retrieval or action failure, and nodding for error conditions or confusions. Sleep and wake agent modes were animated with vanishing and re-appearing agent animation sequences: the agent in sleep mode was represented by a transparent ghost image on the screen which transforms back to the normal image when awakened by the appropriate attention command (e.g., wake up chimp). There were several idle sequences implemented which would automatically put the agent in sleep mode should there be no user action within a specified time out period.

Informal Evaluation: An informal evaluation of the prototype was carried out as a part of the iterative design process. A total of eight children, four girls and four boys ages eight to 14 years, tested the system. Since the system design changed during the course of evaluation (and software fixes were made), only qualitative evaluation results are available. It should, however, be noted that the lessons learnt with this prototyping contributed toward the design of other multimodal mixed-initiative systems that were formally evaluated and reported elsewhere [15], [18]. During the evaluation, the child user was first briefed about the application. A scenario was provided for exercising the features of the communication agent while interaction with the spelling agent part was kept open. The entire interaction with the system was video taped; user's feedback was sought at the end of the experiment.

Overall the prototype received positive feedback from the users especially for the animated agent and the speech interface. Children reported enjoying the naturalness and flexibility of the interface and communicating with the animated agent. In these experiments, we found that the children tended to switch modalities from voice to mouse clicks either when there was repeated ASR errors or when there was a need for dialog disambiguation. The primary criticism of the prototype was the limited range of interactions one could have with the animated agent (there were only about ten animated response sequences) and the lack of understanding of out-of-domain user requests. Specifically, some children were disappointed at the lack of more sophisticated social interaction capabilities for the agent. While some of these research questions are currently being pursued by us and others [2], a formal evaluation of the complete spoken language prototype remains to be done.

V. Summary

Speech input as a component of the multimedia experience fabric is an interaction modality greatly desired by children users. The addition of conversational capability to children's multimedia applications contributes to more natural user interactions and improved user experience. The experiments and results reported in this paper show that it is feasible to build conversational systems for children. The inherent variability in children's speech makes ASR difficult. Speaker normalization and model adaptation were used to improve speech recognition performance. A WoZ experiment in a gaming environment provided data for creating novel language models and understanding strategies for dialog systems. Lessons learned from this study and its concurrent user experience evaluation were leveraged in the design of a prototype multimodal–multimedia application for children. Since the system was designed for children users, heavy emphasis was placed on the interface design. Indeed it was 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) made the application easy to use even for novice users. In addition to the user interface, the prototype served as a test bed for creating a general multimodal system architecture. The main design principle of our system was a modular architecture, where the controller communicates with "stateless" servers via text messages (all state information resides in the template). Seamless integration of all input modalities for our applications is achieved by translating all inputs into text strings that are in turn handled by the spoken language understanding system (or, equivalently, directly generating an equivalent semantic representation corresponding to certain input events). Other features of our system not yet implemented include customizable application content and customizable agent personality. Overall, the prototype represents a successful first effort at building a multimodal system for children with an emphasis on conversational speech. We expect that data from such prototypes will help further conversational human–machine interaction technology.

Acknowledgment

The authors would like to express their sincere appreciation to C. Lazarus, M. Einbinder, and R. Barkan for organizing the "Carmen Sandiego" Wizard of Oz experiments and for collecting and analyzing the user experience data; to S. Chu and H. Wang for their help with the design and implementation of the Agent CHIMP prototype; to S. Lee for the acoustic analysis of the C.I.D. children database; to G. Riccardi for his work in language model adaptation for the "Carmen Sandiego" task; and to R. C. Rose for many useful discussions on speaker normalization and utterance verification.

References

[1] The American Learning Household Survey, "A study of household demand for educational programming, software and technology," Conducted by FIND/SVP: Emerging Technol. Res. Group, Sept. 1995.

[2] S. Arunachalam, D. Gould, E. Andersen, D. Byrd, and S. Narayanan, "Politeness and frustration language in child–machine interactions," in *Proc. Eurospeech*, Aalborg, Denmark, 2001.

[3] D. C. Burnett and M. Fanty, "Rapid unsupervised adaptation to children's speech on a connected-digit task," in *Proc. ICSLP*, Oct. 1996.

[4] P. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Clow, and J. Smith, "The efficiency of multimodal interaction: A case study," in *Proc. ICSLP 98*, Sydney, Australia, 1998.

[5] B. Buntschuh, C. Kamm, G. DiFabbrizio, A. Abella, M. Mohri, S. Narayanan, I. Zeljkovic, J. Wright, S. Marcus, R. D. Sharp, R. Duncan, and J. Wilpon, "VPQ: A spoken language interface to large scale directory information," in *Proc. ICSLP*, Sydney, Australia, 1998, pp. 2863–2867.

[6] S. Eguchi and I. J. Hirsh, "Development of speech sounds in children," *Acta. Otolaryng.*, Suppl. 257, 1969.

[7] G. DiFabrizzio, P. Ruscitti, S. Narayanan, and C. Kamm, "Extending computer telephony and IP telephony standards for voice-enabled services in a multi-modal user interface environment," in *Proc. Interactive Dialogue Multi-Modal Systems*, Kloster Irsee, Germany, June 1999, pp. 9–12.

[8] U. U. G. Goldstein, "An articulatory model for the vocal tracts of growing children," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, MA, 1980.

[9] A. Gorin, G. Riccardi, and J. Wright, "How may I help you?," *Speech Commun.*, vol. 23, pp. 113–127, 1997.

[10] Jupiter Communications Research, Aug. 11, 1998.

[11] R. D. Kent, "Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies," *J. Speech Hear. Res.*, vol. 19, pp. 421–447, 1976.

[12] L. Lamel *et al.*, "The LIMSI RailTel system: Field trial of a telephone service for rail travel information," *Speech Commun.*, vol. 23, pp. 67–82, 1997.

[13] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. ICASSP*, May 1996, pp. 353–356.

[14] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *J. Acoust. Soc. Amer.*, vol. 105, pp. 1455–1468, Mar. 1999.

[15] E. Levin, S. Narayanan, R. Pieraccini, K. Biatov, E. Bocchieri, G. Di Fabbrizio, W. Eckert, S. Lee, A. Pokrovsky, M. Rahim, P. Ruscitti, and M. Walker, "The AT&T-DARPA communicator mixed-initiative spoken dialog system," in *Proc. Int. Conf. Spoken Language Processing*, Beijing, China, 2000, pp. 122–125.

[16] J. Mostow, A. G. Hauptmann, and S. F. Roth, "Demonstration of a reading coach that listens," in *Proc. ACM Symp. User Interface Software Technology*, 1995, pp. 77–78.

[17] "Leisure time study," MTV Network, Aug. 11, 1998.

[18] S. Narayanan, G. Di Fabbrizio, C. Kamm, J. Hubbell, B. Buntschuh, P. Ruscitti, and J. Wright, "Effects of dialog initiative and multi-modal presentation strategies on large directory information access," in *Proc. Int. Conf. Spoken Language Processing*, Beijing, China, 2000, pp. 636–639.

[19] "Teenagers and technology," *Newsweek*, p. 86, Apr. 28, 1997.

[20] "Online attitude and usage study," *Nickelodeon*, Aug. 11, 1998.

[21] R. Pieraccini and E. Levin, "A spontaneous-speech understanding system for database query applications," in *Proc. ESCA Workshop on Spoken Dialogue Systems—Theories and Applications*, 1995.

[22] R. Pieraccini, E. Levin, and W. Eckert, "AMICA: The AT&T mixed initiative conversational architecture," in *Proc. Eurospeech*, Rhodes, Greece, Sept. 1997.

[23] A. Potamianos *et al.*, "Design principles and tools for multimodal dialog systems," in *Proc. ESCA Workshop Interact. Dialog. Multi-Modal Syst.*, Kloster Irsee, Germany, June 1999.

[24] A. Potamianos and S. Narayanan, "Spoken dialog systems for children," in *Proc. ICASSP*, Seattle, WA, May 1998, pp. 197–200.

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

[26] A. Potamianos, G. Riccardi, and S. Narayanan, "Categorical understanding using statistical N-gram models," in *Proc. Eurospeech*, Budapest, Hungary, Sept. 1999, pp. 2027–2030.

[27] A. Potamianos and R. C. Rose, "On combining frequency warping and spectral shaping in HMM-based speech recognition," in *Proc. ICASSP*, Apr. 1997.

[28] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[29] G. Riccardi, A. Potamianos, and S. Narayanan, "Language model adaptation for spoken language systems," in *Proc. ICSLP 98*, Sydney, Australia, 1998, pp. 2327–2330.

[30] 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 *Proc. ICSLP*, Philadelphia, PA, Oct. 1996.

[31] R. Sharma, V. Pavlovic, and T. Huang, "Toward multimodal human computer interface," *Proc. IEEE*, vol. 86, pp. 853–869, May 1998.

[32] E. F. Strommen and F. S. Frome, "Talking back to big bird: Preschool users and a simple speech recognition system," *Educ. Technol. Res. Develop.*, vol. 41, pp. 5–16, 1993.

[33] T. Takezawa and T. Morimoto, "A multimodal-input multimedia-output guidance system: MMGS," in *Proc. ICSLP '98*, Sydney, Australia, 1998.

[34] J. G. Wilpon and C. N. Jacobsen, "A study of automatic speech recognition for children and the elderly," in *Proc. ICASSP*, May 1996, pp. 349–352.

[35] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington, "Jupiter: A telephone-based conversational interface for weather information," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 85–96, Jan. 2000.

Shrikanth Narayanan (M'87–SM'02) received the M.S., Engineer, and Ph.D. degrees, all in electrical engineering, from the University of California, Los Angeles, in 1990, 1992, and 1995, respectively.

From 1995 to 2000 he was with AT&T Labs-Research, Florham Park, NJ (formerly AT&T Bell Labs, Murray Hill, NJ), first as Senior Member, and later as Principal Member of Technical Staff. He is currently an Assistant Professor at the Signal and Image Processing Institute, Electrical Engineering Department, University of Southern California (USC), Los Angeles. He is also a member of the Integrated Media Systems Center, an NSF Engineering Research Center and the faculty in Linguistics at USC. His research interests include signal processing and systems modeling with an emphasis on speech and language processing applications. He is an author or co-author of more than 60 publications and three U.S. patents.

Dr. Narayanan is an Associate Editor of the IEEE TRANSACTIONS OF SPEECH AND AUDIO PROCESSING and serves on the Speech Communication technical committee of the Acoustical Society of America. He is member of Tau Beta Pi and Eta Kappa Nu.

Alexandros Potamianos (M'92) received the Diploma degree in electrical and computer engineering from the National Technical University of Athens, Athens, Greece, in 1990. He received the M.S. and Ph.D. degrees in engineering sciences from Harvard University, Cambridge, MA, in 1991 and 1995, respectively.

From 1991 to June 1993, he was a Research Assistant with the Harvard Robotics Lab, Harvard University, Cambridge, MA. From 1993 to 1995, he was a Research Assistant with the Digital Signal Processing Lab, Georgia Institute of Technology, Atlanta. From 1995 to 1999, he was a Senior Technical Staff Member with the Speech and Image Processing Lab, AT&T Shannon Labs, Florham Park, NJ. In February 1999, he joined the Multimedia Communications Lab, Bell Labs, Lucent Technologies, Murray Hill, NJ. He is also an Adjunct Assistant Professor with the Department of Electrical Engineering, Columbia University, New York. His current research interests include speech processing, analysis, synthesis and recognition, dialog, and multimodal systems, nonlinear signal processing, natural language understanding, artificial intelligence, and multimodal child–computer interaction. He has authored or co-authored over 30 papers in professional journals and conferences. He holds three U.S. patents.

Dr. Potamianos has been a member of the IEEE Signal Processing Society since 1992 and he is currently a member of the IEEE Speech Technical Committee.