Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 24 (2010) 232–256
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Towards modeling user behavior in interactions mediated through an automated bidirectional speech translation system

JongHo Shin, Panayiotis G. Georgiou *, Shrikanth Narayanan

Viterbi School of Engineering, University of Southern California, 3740 McClintock Av., EEB400, Los Angeles, CA 90089-2564, United States

Received 27 March 2008; received in revised form 12 February 2009; accepted 27 April 2009
Available online 22 May 2009

Abstract

This paper addresses modeling user behavior in interactions between two people who do not share a common spoken language and communicate with the aid of an automated bidirectional speech translation system. These interaction settings are complex. The translation machine attempts to bridge the language gap by mediating the verbal communication, noting however that the technology may not be always perfect. In a step toward understanding user behavior in this mediated communication scenario, usability data from doctor–patient dialogs involving a two way English–Persian speech translation system are analyzed. We specifically consider user behavior in light of potential uncertainty in the communication between the interlocutors. We analyze the Retry (*Repeat and Rephrase*) versus Accept behaviors in the mediated verbal channel and as a result identify three user types – *Accommodating*, *Normal* and *Picky*, and propose a dynamic Bayesian network model of user behavior. To validate the model, we performed offline and online experiments. The experimental results using offline data show that correct user type is clearly identified as a user keeps his/her consistent behavior in a given interaction condition. In the online experiment, agent feedback was presented to users according to the user types. We show high user satisfaction and interaction efficiency in the analysis of user interview, video data, questionnaire and log data.

Keywords: Speech to speech translation; User modeling; Bayesian reasoning; Spoken dialog systems; Agent; Feedback; Objective and subjective measures; Machine mediated communication; Cross-lingual interactions

1. Introduction

Spoken conversations have been recognized as the primary communication mechanism between humans. With increasing globalization, the need for cross-lingual interactions has become a necessity for a variety of domains including business and travel. As speech and language technologies evolve, we can envision intelligent speech-enabled systems mediating dialogs between people who do not share a language, through automated speech to speech translation. Significant progress is being made in this direction by several research

* Corresponding author. Tel.: +1 213 740 4654; fax: +1 213 740 4651.
E-mail addresses: jonghosh@usc.edu (J. Shin), georgiou@sipi.usc.edu (P.G. Georgiou), shri@sipi.usc.edu (S. Narayanan).

0885-2308/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2009.04.008

institutions (Narayanan et al., 2003; Zhou et al., 2003; Precoda and Podesva, 2003; Black et al., 2002). The goal of such systems is to be truly cognizant of the interaction, intelligent and performing as a communication aide, beyond serving as a mere message conduit.

Drawing parallels with advances in human–machine spoken dialog systems, we can see that incorporating intelligence into a spoken language based communication mediation system requires, among other things, careful user modeling in conjunction with an effective dialog management. In general, user modeling in systems design has been attempted at different levels and using a variety of approaches. Rich (1999) has proposed a 3-dimensional space of models: individual user versus generic; explicitly defined versus learned; and long-term user characteristics based versus short-term user behavior based models. In Table 1 the three axes of these descriptors relate to the size of the population the model describes, the fashion in which the model is created and also the temporal scale the model is attempting to characterize.

While there has been a fair amount of excellent user modeling work in the context of human–machine spoken dialogs including user simulation (Eckert et al., 1997; Georgila et al., 2005), reasoning about a user's goal or intention (Horvitz et al., 1998), user expertise modeling (Komatani et al., 2003), and evaluation techniques (Litman and Pan, 1999), relatively little effort has been devoted in this regard on machine mediated human–human cross-lingual dialogs, the topic of this paper. The motivation stems from the need for informing designs of speech translation systems for their increased effectiveness and usability as communication aids.

Construction of a user model based on the desired user features, however, can be a daunting task. Generally, two approaches – "Profiling modeling" and "Statistical modeling" – are widely used in this regards. The profile acquired from a user can be used for generating an appropriate system response, such as personalized search (Pitkow et al., 2002), or in providing appropriate help to the user when needed (Horvitz et al., 1998; Bauer, 1999; Yan and Selker, 2000). In this present work, we adopt the second approach, where predictive statistical user models are derived from usage data. It is considered a powerful approach to model user behavior (Zukerman and Albrech, 2001), and its effectiveness has been demonstrated by previous research (Komatani et al., 2003; Kuenzer et al., 2001). We specifically propose a Bayesian network user model for our analysis to exploit its effective reasoning capabilities under uncertain situations.

In order to study user modeling issues in speech-to-speech translation systems, we consider two separate but mutually dependent channels – the human–machine–human (machine mediated) and the direct human-to-human (interpersonal) channels. The verbal communication is handled through the machine, and effects of uncertainty and errors in the machine processing can be expected to be predominantly manifested in the verbal behavior of the user. On the other hand, the interpersonal channel is characterized by direct gestural non-verbal exchanges (such as head nods) as well as indirect verbal means (such as through adaptation to each others speaking styles). Our analysis in this paper is restricted to aspects of the verbal behavior in these channels. The rest of the paper is organized as follows. After a description of the speech-to-speech system used in this study for doctor–patient interactions and the corresponding data in Section 2, in Section 3 we analyze and model user behavior in the mediated channel under potential uncertainty by focusing on the "Retry"(*Repeat/Rephrase*) behavior. We describe a dynamic Bayesian model to predict such behavior and evaluate its performance in offline data. In Section 4, we present an online experiment with agent feedback and report the results. Finally, conclusions and a description of future work plans are given in Section 5.

2. System and dataset

2.1. A two-way speech translation system with a push-to-talk interface

The system used for the study of this paper is a speech-to-speech translation device that facilitates two way spoken interactions between an English speaking doctor and a Persian (Farsi) speaking patient (Narayanan

Table 1
User model dimensions based on the knowledge about people according to Rich (1999).

Dim. 1	A single, canonical user	A group, collection of users
Dim. 2	Specified by the system designer	Inferred by the system
Dim. 3	Long term	Short term

et al., 2003). An excerpt from a transcribed conversation is presented in Table 2. This version of the system uses a push-to-talk modality to initiate a speaking turn which has its advantages and limitations. The push-to-talk interface minimizes recognition and translation errors since users can verify concepts before executing the final decision for "speaking out" the translation but has the disadvantage of creating less spontaneous and less natural interactions.

Furthermore, the goal of the system is to facilitate a task oriented rather than a free-form social interaction between the two participants. Specifically, the domain of usage of the system under study is task-specific (or goal-oriented) interaction between a doctor and a patient. Within this context, the system design strives to achieve not only optimal technology performance, such as of automatic speech recognition and translation, but also maximal user satisfaction. Prior work has clearly shown that user satisfaction is one of the most important efficacy metrics of medical domain interactions (Hall et al., 1988; Roter and Hall, 1989).

A functional block diagram of the system used in the present study and its data flow are shown in Fig. 1. The user's spoken utterance is converted into textual form by an automatic speech recognizer (ASR) in the appropriate language of the speaker (English for the doctor and Farsi for the patient in this case) and further processed by two parallel mechanisms: one by a phrase-based statistical Machine Translation (MT) module that translates the text from one language to another and the other by a statistical classifier which attempts to categorize the utterance into one of several predetermined "concept" categories. The Dialog Management (DM) module interacts with the MT/classifier and the GUI and TTS modules to deliver the data to the user. In the system of this study, the visual output provided by the GUI is made available only to the (English-speaking) doctor, who is assumed to have the primary control of the interaction.

Table 2
Excerpt from a conversation between a doctor-role participant and a patient-role participant who used Transonics.

(1)	Doctor said:	WHAT BRINGS YOU TO THE CLINIC?
	Patient said:	dAStm kh dl drd myknm (I HAD THE ABDOMINAL PAIN)
(2)	Doctor said:	DOES YOUR CHEST HURT?
	Patient said:	bly (YES)
(3)	Doctor said:	DOES YOUR ARM HURT?
	Patient said:	VyA drd fqT AynJAst (DOES THE PAIN STAY RIGHT HERE)

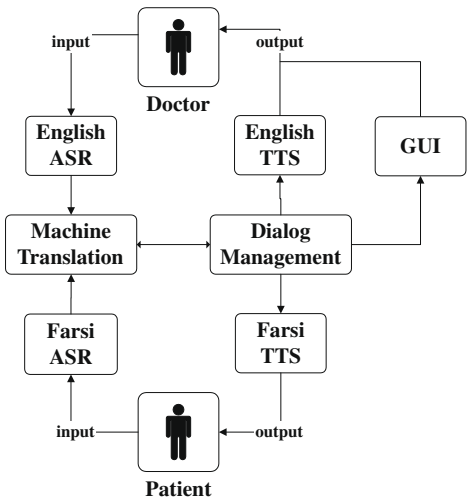

Fig. 1. Simplified data flow diagram of our two way speech translation system for doctor–patient interactions. English and Farsi Automatic Speech Recognition (ASR) models get the input from users (doctor and patient, respectively) while the Machine Translation (MT) module is responsible for automatic translation and classification of the input. The Dialog Manager (DM) manages the interaction and communicates the translated results to a graphical user interface (GUI) and a text to speech (TTS) synthesizer (in English and Farsi as appropriate).

To better understand the translation device operation and the associated issues, we can identify three distinct operations in the process that can introduce uncertainty into the communication chain. The first inherently lossy operation is the conversion from speech into a textual transcription of the spoken utterance through statistical pattern recognition (ASR) i.e., often the transcript may not accurately represent what the user spoke, characterized by deletion/insertion/substitution of words. The second one is the translation. We have two concurrent statistical approaches to this step (statistical machine translation and an utterance concept classifier) that represent a lossy text mapping. The third stage is the conversion of the target language transcript from text to audio by synthesizing the speech through text-to-speech (TTS) synthesis, which can be lossy due to several reasons including due to operating on the noisy output from the ASR and translators. All these potential information losses can impact the communication between the participants.

By design, the interface control of our experimental system was asymmetric in the sense that the (English-speaking) doctor had exclusive control over the interface, and access to the GUI, while the (Farsi-speaking) patient did not. This was to allow even untrained and non-educated patients access to the system. The system allows for the doctor to decide whether to transmit one of the several alternate hypotheses offered by the system to the patient or reject all of them (repeat or rephrase). Some of the options provided to the doctor can be seen in Fig. 3 and the hypotheses belong to one of two classes:

1. The first is the English transcription of what the machine thinks the user said. The machine does not provide a translation on the screen (presumably it would not be useful for the doctor who does not know Persian) but a statistical phrase based translation would be provided to the patient if the doctor chooses this option. However, such statistical machine translation *cannot* guarantee accurate translation of the displayed text. This option mainly allows the user to detect errors from the ASR stage of the translation process, and thereby reduce the risk of error during the translation.
2. The second category of options takes the recognized transcript (output of ASR stage) and maps it into one of several pre-determined concept categories. These categories were manually specified and for this domain there were about 1200 concepts. This mapping operation from text to concept is also lossy, but unlike the first hypothesis, since these concept categories are pre-programmed in the system, a back-translation (canonical form) in the language the doctor understands can be displayed for the doctor's choosing. This means that what the doctor sees on the screen already includes any errors likely made by both the ASR and

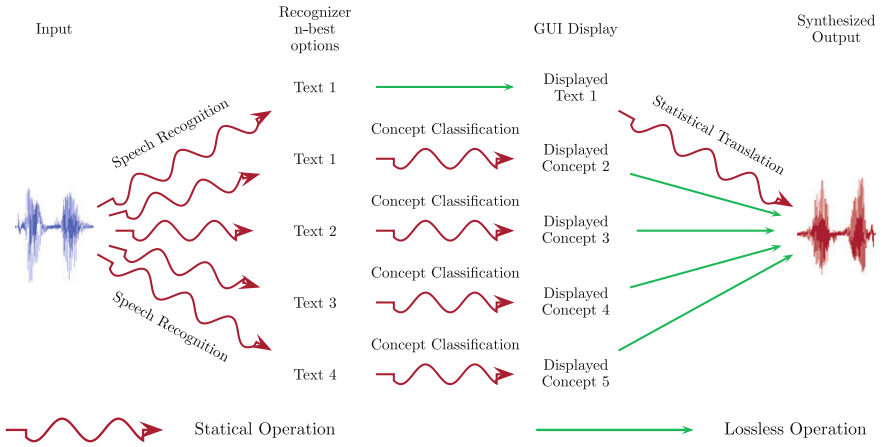

Fig. 2. The internal procedure of generating speech translation hypotheses in our system. Two parallel mechanisms are implemented. In the first one, the topmost recognition candidate i.e., the first-best choice of the ASR – that has already gone through a lossy speech to text mapping process – will go through another lossy operation – the statistical translation. In the second one, that utilizes an utterance classifier, the top four recognized candidates from the ASR (the so called four-best results) are mapped into conceptual classes, also a lossy operation, but the canonical form result – after both lossy operations – is the one displayed on the screen for the doctor's choosing.

Fig. 3. Transonics system screen GUI. After speaking, the user (doctor) can choose one of several hypotheses presented on the GUI.

Table 3
DARPA evaluation on medical domain for the speech translation system of this paper. Component and Concept measures are: ASR word error rate (lower is better), SMT BLEU score (higher is better) with the clean text transcript input or with the ASR output as an input.

DARPA evaluation results		
	English	Persian
ASR WER	11.5%	13.4%
	English to Persian	Persian to English
IBM BLEU (text)	0.31	0.29
IBM BLEU (ASR)	0.27	0.24
Overall concept transfer		78%

translation steps, and that the translation the patient will hear will be lexically identical to the hypothesis displayed on the screen. Fig. 2 depicts these procedures conceptually. It is clear that if one of the canonical sentences is satisfactory from a concept transfer perspective, it should be the best choice for the user since these guarantee accurate translation.

Users of the device were encouraged to employ the second category of options (labeled on the GUI: "I can definitely translate these") if these options were deemed valid representations of their utterances, rather than the first option (labeled on the GUI: "I can try to translate this"). For example, in Fig. 3 when the doctor says "You have fever?" the device can try to translate the ASR text output "You have fever" or it can definitely say "Do you have a fever?", the surface form for a concept category related to "fever-inquiry".

The monolingual patients on the other hand are assumed to be untrained in using the system – and to ensure uniform results in the experiments described in this paper – are not allowed to see the screen. The system decides, based on confidence scores of automatic utterance to concept classification, whether their utterance is close enough to a particular concept class. The cluster-normalized form concept will be transferred to the doctor if deemed a good translation, or instead a direct potentially noisy statistical translation of the text will be provided. Most of the time an incorrect transfer can be detected by the doctor due to the lack of coherence with the discourse of the interaction. The Persian patient can also choose to request, verbally or through gestures, repetitions or repairs if they so chose. Note that an experienced doctor, in the case of receiving information that does not match the discourse can assume that he needs to do error control by rejecting the solution provided by the system (and repeat/rephrase).

In terms of component level performance of the system used in the present study, the ASR word error rate, the concept transfer rate and the IBM BLEU translation score are given in Table 3. These results stem from

Table 4
Table shows a simplified portion of the data log acquired automatically by running the Transonics speech translation system. There are system routing tags (FADT, FDMT, FMDT, FDGT, FDGC, FGDT – F: flow, A: audio server, D: dialog management, M: machine translation, G: graphical user interface, T: text, and C: control) indicating the data flow from/to on the left side and the data being processed on the right side. Actual data are in the content column. Additional information logged, not shown for simplicity, include time stamps, utterance sequence, confidence and class numbers.

System routing tag	Content
FADT	YOU HAVE OTHER MEDICAL PROBLEMS | DO YOU HAVE OTHER MEDICAL PROBLEMS
FDMT	YOU HAVE OTHER MEDICAL PROBLEMS
FMDT	SmA mSkl pzSky dygry dAryd | YOU HAVE OTHER MEDICAL PROBLEMS
FDGT	YOU HAVE OTHER MEDICAL PROBLEMS
FDMT	DO YOU HAVE OTHER MEDICAL PROBLEMS
FMDT	VyA hyC mSkl pzSky dAryd | DO YOU HAVE ANY MEDICAL PROBLEMS
FDGT	DO YOU HAVE ANY MEDICAL PROBLEMS
FDGC	Shown All Options
FGDT	Choice $*$ 1

the evaluation done under the DARPA Babylon program. The overall concept transfer rate of the system is 78% – this denotes how many of the key concepts (such as symptom descriptions) were correctly transferred overall in both languages according to human observers for the 15 sessions examined in this paper. Also, in Table 3 the word error rate (WER[1]) and the IBM BLEU[2] scores are provided.

2.2. Data-set

The data analyzed for the user modeling purpose are from 15 interactions between doctors and standardized patient actors (details in Narayanan et al. (2003)). Both the doctors and patients are monolingual and, in addition, acoustic masking was in place to ensure translations are only being transferred through the device. The spoken interactions were logged by the system and also transcribed manually. Automatic logs contain recognized utterances (hypotheses) of the ASR, all translated hypothesis from the translation component (both SMT and classified concepts). These come with the confidence levels and the system level information. A simplified portion of a data log is presented in Table 4.

Automatic tagging of the retry behavior was made possible through system logs, and the speech recognition WER scores were acquired by comparing automatically recognized utterances and their human generated transcriptions. It may be interesting to note some relevant information regarding the data characteristics. The average number of turns (each turn is assumed to be a doctor or a patient utterance) in a conversational dialog is 30.13, with a slightly higher number (33.46) for the doctor than for the patient (26.8) with standard deviation of 8.7 and 10.6 respectively. The longest utterance was 13 words long for both the doctor and patient side, while on average utterance length was 4.45 and 2.42 words for the doctor and patient, respectively. The shorter average utterance length of the patient reflects the fact that a significantly large number of their answers were short, such as yes/no answers. The total time for the whole data set is 4 h.

Because of the dynamics created by the push-to-talk interface (managed by only the doctor), the doctor-side data contains abundant information we can utilize to model user behavior in the mediated (verbal) channel.

[1] Word Error Rate is the sum of the number of words in error (substitution, deletion and insertion) divided by the number of words in the reference transcription.

[2] In simple terms, the more ways a certain utterance can be translated, the lower will be the maximum possible score, since one translation will be compared with many possibilities. So although the score is on a theoretical scale of $0 \leqslant \text{IBMBLEU} \leqslant 1$, even the best human expert translators can only achieve average ranges of near a half of that.

3. The mediated channel

We refer to the information path between the two participants through the machine as the *Mediated Channel*. In this channel, a user is cognizant of the machine and acts by considering both the response of the system and his own prior actions. Also, the system can detect how a user behaves or what information is going through the channel. In this sense, it can be regarded as similar to a human–machine interaction scenario.

The methods of identifying the user's model from interactions with a device include investigating behavior patterns (Pitschke, 1994; Manavoglu et al., 2003) and stereotypes (Rich, 1979). Following these generally classified assumptions, considerable research efforts have been undertaken covering various topics and systems: Komatani et al. (2003) introduced a general user model with skill level, knowledge level, and degree of urgency in a spoken dialog system, Carberry et al. (1999) modeled user preferences in a natural language consultation system; Conati et al. (2002) proposed how to manage uncertainty in a student model by performing assessments and recognizing plans for a tutoring system; and Prendinger et al. (2005) utilized physiological data for determining affective states for an emotion recognition system. Furthermore, some frameworks have been suggested for rapid and efficient implementation of user models such as in Kobsa and Pohl (1995), Pakucs (2003), and Tsiriga and Virvou (2004).

Error handling mechanism is an important aspect in the design and optimization of a spoken dialog system. As mentioned, earlier the spoken communication channel between a human and a machine is inherently noisy, which can further be exacerbated by user-dependent uncertainty such as due to limited vocabulary or task knowledge. The significance of considering user behavior under problematic conditions in human–machine interaction is demonstrated, for example, by our prior work (Shin et al., 2002) where we highlighted the importance of repeating and rephrasing cues. Similarly, the work of Batliner et al. (2003) utilized the features such as prosody and linguistic behaviors to model and recognize trouble in communications. Detection and modeling of problematic communication conditions helps to prevent and recover from errors effectively.

Specific user behavior patterns can be attributed to specific user types. Similar to the notion of expert/novice users, in this work, we consider the idea of identifying accommodating and non-accommodating ("Picky") user types under problematic interaction situations. The motivation being that distinct interface strategies can be developed for each user type case, furthering the overall performance. Our experimental analysis indicates that for the same average speech recognition WER, one user retried 95% of the time while another user only 65%. For example, we have observed that certain users are more accepting of minor errors in translation and recognition (e.g., function word insertion such as in "And do you have fever?" when they actually spoke "Do you have fever?") while others completely reject such a hypothesis from the machine as not their intended utterance, despite the fact that it conveys for all practical purposes the identical meaning.

We therefore propose modeling users in one of three categories (*Accommodating, Normal and Picky*) based on the analysis of the active participant, the doctor: for example, *Accommodating* users tend to accept more system errors than the other user types. Following which, we train a system that can detect in which category the user belongs based on the user behavior through the interaction history and current utterance features. While devising specific interventions based on the model outcome is not the goal of this paper, we hope that this approach will, however, enable future research in building agents that can appropriately adapt the system according to detected user behaviors similar to what previous studies have demonstrated (Jokinen and Kanto, 2004; Kamm et al., 1998; Komatani et al., 2003).

3.1. Analysis of repeat/rephrase("Retry") behavior

Repeat or rephrase (Retry) was the primary user behavior observed under problematic conditions caused by non-optimal or poor system performance in the Transonics system. In addition to the user type being an important factor in determining the degree of retry actions, the level of speech recognition error was found be an important factor. However, in our *standardized subject*[3] experiments, the difference range of the speech recognition error among users is small; therefore we assume that the user type has a stronger effect on the

[3] The subjects were all native US English speakers, medical professionals and trained equally before using the system.

observed retry behavior. In addition to the small variance in the speech recognition error, we observed that most errors stem from insertions of function words and that keywords are mostly correctly recognized. Typical examples of errors with erroneously inserted words underlined are "A how are you" or "tell me THE about your pain". Other potential contributing factors such as user's emotion, knowledge, gender, physical condition, hastiness, etc. are not considered at this stage but are of interest and will be included in the analysis once larger data sets become available.

3.1.1. Categorizing user types: Accommodating, Normal and Picky

User type is a casting of a user along several categories; it can be based on demographic information, such as *Gender* or *Age* or a heuristic category such as *Expertise* or *Knowledge level*. We consider, in this paper, the degree of user accommodation to spoken language processing errors as the criterion to decide a user type. The use of such heuristic domain-specific criteria has been prevalent in user modeling research. For instance, in Komatani et al. (2003), user skill level is defined by the maximum number of slots filled by utterances and in Kobsa (1990) and Conati et al. (2002), knowledge level is decided based on correct answers to the domain questions. In most cases, heuristic methods are used for user type classification even though those may not always be too accurate. For example, if we assume that knowledge level is judged by the number of correct answers to system questions, this is usually a good metric, but it is not a perfect one since the user may give wrong answers on purpose to trick the system, may be tired and not pay enough attention, or may not be motivated enough to devote the necessary attention.

For our off-line model, we cluster user types based on the total number of retries of each user. We assume that accepting different ranges in WER depends significantly on the user type, as conceptualized in Fig. 4, and hence we define

- *Accommodating*: users tend to accept highly erroneous transcriptions compared to other users.
- *Normal*: users accept some degree of errors.
- *Picky*: users tend to reject all but the most exact transcriptions, thus being very strict in what they accepted for translation.

Based on data from the 15 sessions analyzed in this work, we clustered the users with the k-means algorithm into the three classes as shown in Fig. 5. Note that one could argue in favor of fewer or more quantization steps along the accommodation axis. Such decisions depend more on the action to be taken upon classification and the available data for the analysis.

From the clustering results, 7 (47%) users present themselves as Accommodating, 5 (33%) as Normal and 3 (20%) as Picky. The users tend to *retry* at different degrees: *Accommodating* 19.3%, *Normal* 31.3%, and *Picky*: 40.7%. The average WER rate across *all* the utterances, however, does not vary significantly and stands at 35.9, 43.8 and 38.7 for *Accommodating*, *Normal* and *Picky*, respectively. Hence we did not employ WER as a feature for the clustering of user types. Note that although the average WER is relatively constant from user to user, the error that users consider acceptable is not, as demonstrated by the variable degree of retries.

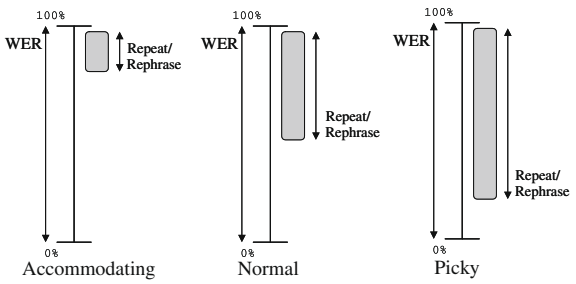

Fig. 4. The *Accommodating* user tends to "Retry" significantly less than the other users while the *Picky* user tries significantly more. A user in between these extremes is defined to be a *Normal* user. WER is the speech recognition Word Error Rate and the above graph conceptually demonstrates the ranges of WER for which each user type tends to "Retry."

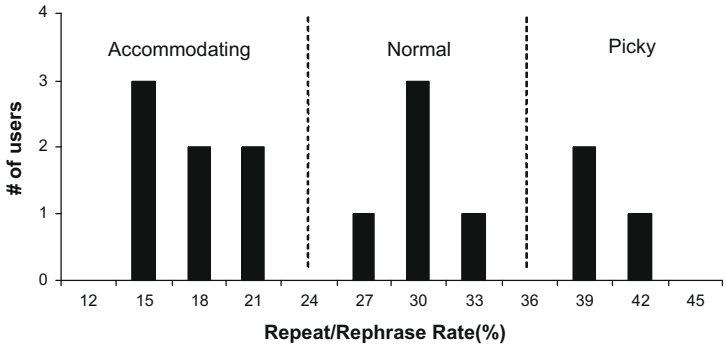

Fig. 5. The quantized retry rate over 15 interaction sessions on the doctor side. The criteria (average retry rate) based on the data analysis led us to categorize the users into 3 types: Accommodating, Normal, and Picky.

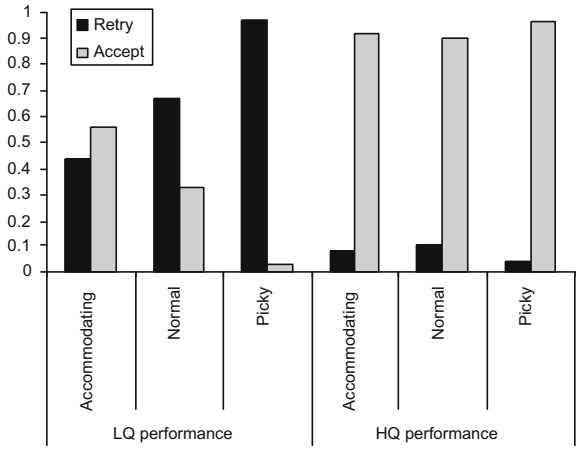

Fig. 6. Conditional Probability Table (CPT) over user behaviors(discrete) – "Retry" and "Accept". Each user type is represented numerically with regard to Low Quality (LQ) and High Quality (HQ) system performance (recognition error rate). The *Y*-axis represents the probability of user behavior conditioned on user type and system performance.

Assuming a certain threshold separating the high-quality (HQ) speech recognition performance from a low-quality (LQ) performance (a detailed discussion of how the two regions of performance can be decided is provided in the next section, Section 3.1.3), we empirically acquired the Conditional Probability Table (CPT) over all the 15 interactions as shown in Fig. 6. We can clearly see the difference in user accommodation when operating in the LQ region.

When the condition represents relatively high system performance (HQ performance), "Accept" behavior dominates covering over 90% in most cases, and allowing us very small amounts of data for observing the "Retry" behavior. With this data in HQ performance, we cannot tell differences in the user "Retry" behavior statistically.

3.1.2. User behavior model with the Transonics system

Since in our analysis we observed that the system error alone cannot account for the large variability in user actions, we hypothesize that the user type combined with the system error under problematic conditions affects the retry behavior. The following conditions are assumed: (1) The system is stationary and the performance is shown in Table 3; (2) The subjects are native speakers (US English) and user performance is consistent in terms of machine recognition (no acoustic/lexical mismatch issues in speech recognition); (3) Domain knowledge of subjects is the same (all medical professionals); (4) Skill differences and adaptation levels of subjects

are minimized based on the given environment (trained with equal time and materials and provided the same experimental environment for equal time).

3.1.3. Threshold of high/low quality system performance

Another important issue we need to deal with is the threshold of average acceptable WER for each user. This is a complex issue that is related to each user's personal preferences and traits. We empirically approached this problem with the relative WER average based on retry and accept behaviors across all other users. We assume that a user retries if the system performance falls below a threshold, thus we clustered the per-utterance WER into two groups: the group of accepted utterances and the group of the utterances that are rejected. The low quality (LQ)/high quality (HQ) performance threshold is the separating point of the two clusters at a WER of 56% for the data of these 15 interactions. This implies that there is a high probability of a retry if the WER increases above 56%. For training and testing purposes, the threshold is acquired in a *n*-fold cross-validation from 14 interactions and tested on the remaining 1 interaction. Note that although the threshold WER may seem to imply a very low accuracy for allowing a concept transfer, the classifier frequently may allow accurate concept transfer with WER much higher than that if a keyword has been recognized correctly and the classification gave at least one option which is valid. For example: "Are you having a headache now?" will have a classifier top choice of "Do you have a headache?" even if only the word "headache" has been correctly recognized by the ASR.

3.2. A dynamic Bayesian network user behavior model

A dynamic Bayesian network is a promising representation for modeling the inter-causal relationships of "Retry" behavior with temporal information. The promise of this model has been highlighted in the user modeling field across various applications. The Lumiere project (Horvitz et al., 1998) utilized Bayesian models for capturing the uncertain relationships between the goals and needs of a user. Conati et al. (2002) used Bayesian network to model a student for an automated tutoring system which assesses the knowledge, recognizes plans and predicts actions of each student. Recently, Grawemeyer and Cox, 2005 modeled users' information display preferences by using Bayesian reasoning. Also, the theoretical benefits in its performance and extensibility as a classifier have been thoroughly described in Friedman et al. (1997).

In spite of their remarkable power and potential to address inferential processes, there are some inherent limitations and liabilities to Bayesian networks. First, a Bayesian network cannot represent every possible situation (uncertainties and dependencies) and it takes a long time to choose necessary nodes for the network. Second, the prior knowledge (probability) of each node of the network may be biased, depending on the measurement approach, and this may distort the network and can generate unreliable response to a user. For example, in Horvitz et al. (1998), experts constructed Bayesian models for several applications, tasks and sub-tasks by doing user studies; however, that assumes sufficient and representative coverage of user activities in the observed data.

The details of the proposed DBN implementation are presented in the following sections and general user type prediction algorithm is given in the Table 5.

Table 5
User type inference algorithm computes the probability of user types, *Accommodating*, *Normal* and *Picky* respectively. Each user type is predicted by Bayesian reasoning and updated until one of them becomes believable.

Input: User behavior("Retry" or "Accept") and HQ/LQ recognition information
Output: The most believable user type
Initial: User types with the same probability
Step 1: The probability of each user type is given by the Bayesian reasoning
Step 2: Update the prior of each user type
Step 3: Check whether the belief of the highest user type probability is enough
Step 4: If it is not enough to be believed, go to the Step 1
Return A user type with the highest probability

In this analysis the variables of user behavior (retry/accept) and the system feature, the utterance confidence score (or for off-line processing WER), are the observed variables and the user type is the unknown variable. In the design phase, the network is built by learning parameter values and interrelations of user type and observed variables.

The user type is assumed to be constant, despite the fact that some user characteristics may vary during the course of an interaction. For example, talkative people may be more reserved in communicating when depressed, tired or under stress. A person who is in general sensitive to any kind of system errors can ignore those when he/she is busy. In addition, we often observe that users take time to exhibit their steady state behavior due to an initial adaptation to the other entity, be that a human or a system. It is assumed that the executed behavior and observed feature value are the best representatives for the user type at each time, and the model with these variables is extended dynamically with the temporal information.

We are operating under the assumption that information about the user type could help in altering the system strategy. In addition, this strategy enhances the experience of the user–machine interaction similar to the use of expertise model developed in previous efforts and employed in efficient system strategy design (Kamm et al., 1998; Komatani et al., 2003).

3.2.1. A model of user behavior over a single iteration

We quantize the variables of user type (UT), behavior (B), and system accuracy (F) and these satisfy:

$$\sum_{i=1}^{n} P(UT = ut_i) = 1$$
$$\sum_{i=1}^{m} P(B = b_i) = 1$$
$$\sum_{i=1}^{k} P(F = f_i) = 1 \quad (1)$$

where we chose $n = 3$ discrete levels for the user type, $m = 2$ for behavior and $k = 2$ for the WER. Note that we represent variables by an upper-case letters (e.g., UT, B, F) and its values by that same letter in lower case (e.g., ut, b, f).

The Bayesian network in Fig. 7 shows the complete directed graphical model (static) with the relations among a specific behavior, user type, and features (including unknown features).

Multiple features can exist, and each can have a different effect on the user behavior. Prior work has demonstrated that fewer features are better for improved accuracy/performance (Dash and Liu, xxxx), particularly in small data-sets. Also, unimportant features can be eliminated by utilizing probabilistic measures related to the features (Sheinvald et al., 1990). In the design of the suggested Bayesian model, we chose to incorporate only one feature due to the small amount of data: the quantized (HQ/LQ) WER variable is incorporated with an independent user type variable.

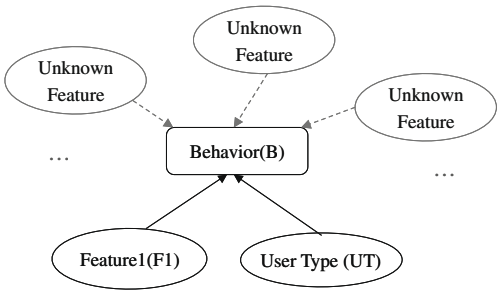

Fig. 7. A generic directed graphical model; the Bayesian network represents the relation in which a user behavior (B) is influenced by a user type (UT) and a feature (F1). There may be unknown features such as emotions and skill level but only one feature is considered for the suggested model.

Based on this general procedure, an actual sequence of stepwise conditional probabilities is formed as in Eq. (2) with the random variables of parents (UT and F) and a child(B). In the user behavior model, we assume that there is no relationship between user type and feature

$$P(B, UT, F) = P(B|UT)P(UT)P(B|F)P(F)/P(B) \tag{2}$$

where $B =$ user behavior, $UT =$ user type, $F =$ feature.

Once the network structure is defined and the conditional probability is decomposed, the quantization of the data in the chosen levels needs to take place. In the suggested model, we have 2 discrete levels for user behavior (retry/accept) and system performance (HQ/LQ) and three user types (*Accommodating*, *Normal* and *Picky*). To give a value for each discrete level, we can utilize a domain expert's knowledge or learn it from the data-set. The second method is adopted in this experiment and the values are learned in a n-fold cross-validation from the training data-set (using 14 out of 15 interactions) for testing on 1 interaction allowing for presenting averaged results over a total of 15 experiments for the 15 interactions in the corpus.

3.2.2. A dynamic model – temporal belief reinforcement

In reality, it takes time to grasp an accurate user type by observing user behaviors and factors (features). For example, observing a one-time accommodating behavior of a user is not enough to decide a definite user type while the observation of consistent behavior over time strengthens the belief of the user's type. This idea is formulated as a dynamic Bayesian network (DBN) shown in Fig. 8. The user type transition mechanism from time $t-1$ to t is supported by the Markovian property that the conditional probability of the current user type(t) depends on the previous user type($t-1$) and it includes the history implicitly by this assumption.

During training, we employ the complete interaction to reason on the user type by using the Maximum Likelihood Estimate (MLE) as in Eq. (3):

$$P(B|F, UT) = \frac{P(F, UT, B)}{P(F, UT)} \tag{3}$$

where $UT = \{ut_1 \ldots ut_n\}, B = \{b_1 \ldots b_m\}, F = \{f_1 \ldots f_k\}$.

The prior for the feature, Word Error Rate (WER), is also acquired from the training data, and the prior of the user type is initially set equally distributed and updated dynamically.

In the absence of large amounts of training data, unconstrained identification of the priors of transition probabilities in a data-driven fashion is not feasible. We instead place parametric constraints on the transition probabilities and identify these parameters in a data-driven fashion. The parameters are the probability of:

- Staying in the same type. This probability is expected to be the highest (P_{SameType}).
- Transitioning across adjacent types (Normal to/from Accommodating and Picky) ($P_{\text{WithNormal}}$).

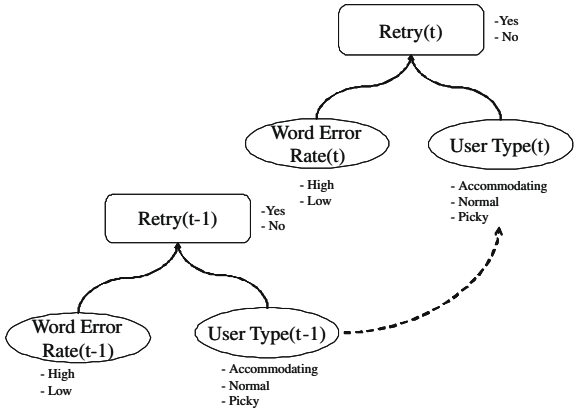

Fig. 8. A dynamic Bayesian network is used to infer a user type over time in the mediated channel. The belief of a user type becomes strengthened as the interaction progresses.

- Transitioning across opposite types (Accommodating to/from Picky). Expected to be the lowest probability (P_{Opposite}).

In addition we define a parameter that reinforces beliefs over time by modifying each of the above probabilities and is defined in terms of the ratio:

$$\mu = \lambda \frac{(\text{Turn number})}{(\text{Total number of turns})} \tag{4}$$

where λ is expected to be a very small number because we want smoothly increase the same-user type transition probabilities over time. Resulting in:

$$\begin{aligned} P_{\text{SameType}}(n) &= P_{\text{SameType}}(n-1) \times (1+\mu) \\ P_{\text{WithNormal}}(n) &= P_{\text{WithNormal}}(n-1) \times \left(1 - \frac{1}{3}\mu\right) \\ P_{\text{Opposite}}(n) &= P_{\text{Opposite}}(n-1) \times \left(1 - \frac{2}{3}\mu\right) \end{aligned} \tag{5}$$

Note that the probabilities are normalized in each turn.

Table 6 presents the values of the parameters. We can also observe that over time the probability of transitioning across opposite types will decay faster than the probability of transitioning across adjacent types.

To infer a user type, the posterior probability of user type conditioned on behavior and feature is computed as in Eq. (6) by applying Bayes' rule:

$$P(UT|B,F) = \eta P(B|UT,F)P(UT) \tag{6}$$

The user type is independent of the observed feature therefore $P(UT) = P(UT|F)$, while $\eta = P(B|F)$ plays the role of a normalizing factor, ensuring that probabilities of user types sum to one.

At each turn, by maximizing the probability of each user type(ut_i) as in Eq. (7), we obtain an estimate of the most probable user type; however the decision is not made until confidence in the belief of user type is significant:

$$\text{argmax}_i \; P(ut_i|B = b_1, F = f_1) = \text{argmax}_i \; P(B = b_1|ut_i, F = f_1)P(ut_i) \tag{7}$$

where $b_1 =$ an evidence of the user behavior, $f_1 =$ an evidence of the feature.

In identifying when a decision on the user's type can be made, we need to consider an acceptable confidence "*Threshold*". This includes two dimensional conditions, when and how to draw a conclusion from the inference. One approach is to decide the final user type when all the available data has been processed (the last state of the DBN) and the evaluation in Section 3.3 is based on this method. An alternative approach is maximum entropy, a good measure that has been utilized in previous work to classify user behaviors (Manavoglu et al., 2003). This may be a more objective and concrete measure of convergence and more appropriate for real-time implementations. As in Fig. 9, we can see the tendency of decreasing entropy for the user type probabilities over all 15 interactions. The entropy decreases as the DBN converges and a lower entropy means that the intra-speaker probabilities of user type are more discriminating. To utilize this mechanism, we could set a certain threshold below which a decision would be made. Otherwise, a user type would be labeled as still unpredictable or not inferable.

Table 6
Values of transition priors. The parametrization allows four variables to represent nine time-varying priors, thus allowing estimation from limited data.

	UT^t_{Acc}	UT^t_{Nor}	UT^t_{Pic}
UT^{t-1}_{Acc}	0.90	0.05	0.05
UT^{t-1}_{Nor}	0.05	0.90	0.05
UT^{t-1}_{Pic}	0.05	0.05	0.90
λ	0.05		

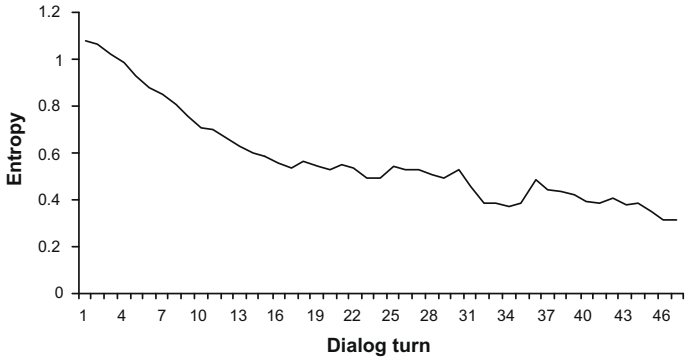

Fig. 9. Entropy of three user types becomes lower as the dialog turn increases. The threshold of deciding the final user type can be set based on this tendency under a dynamic Bayesian reasoning.

3.3. Model validation

We evaluated the automatic identification of the user type by employing the n-fold validation, thus using 14 interactions for training and one for testing, and performing a total of 15 experiments. The goal was to identify user type through the interaction data. Priors were set to be equal (0.33) for the three user types. The classification was successful in 13 out of the 15 dialogs examined by assuming a convergence of the DBN at the end of the available data (method 1, described above). Both errors occurred in identifying the Normal user type, and in both cases it was clear that convergence had not been reached. The DBN was fluctuating between *Normal* and *Picky* in one case and *Normal* and *Accommodating* in the other case. We believe that this may reflect a switching user behavior where users may behave as Picky (if the error is, for example, in a keyword) or as Accommodating (if all the errors are in function words), or it may reflect users who exhibit behavior very close to the user type quantization boundaries.

In the following sections, two representative results of *Picky* and *Normal* user type inference by the suggested DBN model are presented.

3.3.1. Analysis of the Picky user type inference result

Dynamic inference results on an interaction (labeled as *Picky* type) that lasted over 44 turns is depicted in Fig. 10. We can observe that the belief of the *Picky* user type is strengthened over time and is detected early on

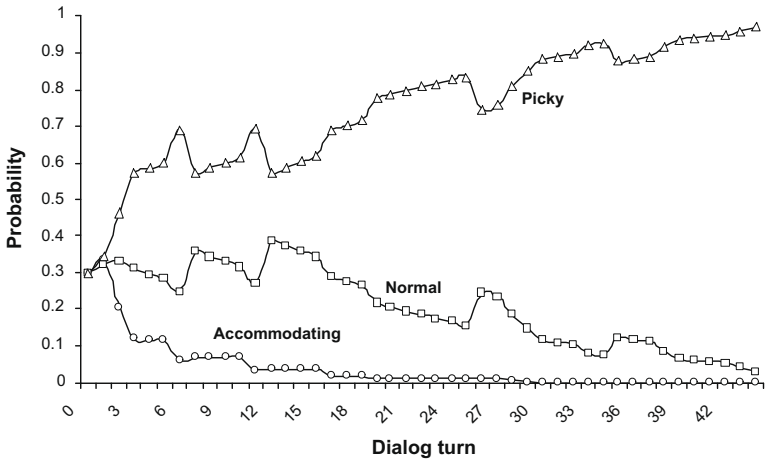

Fig. 10. The belief that the user type is "*Picky*" is strengthened over time in this example data set.

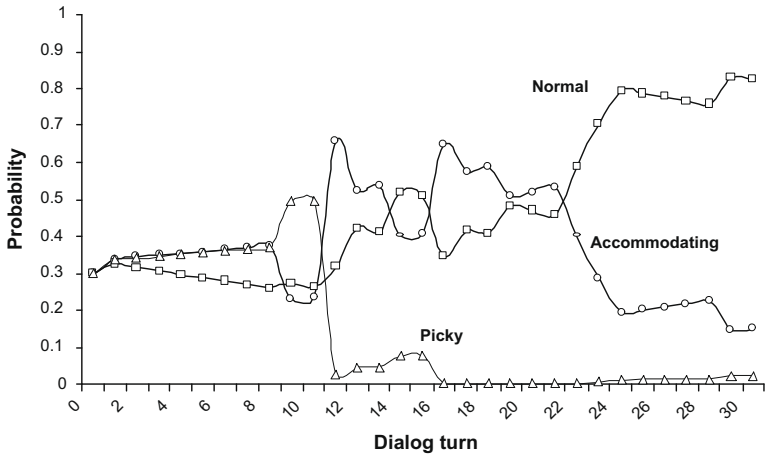

Fig. 11. The belief that the user type is "*Normal*" is strengthened slowly over time.

in the interaction. This implies that a user strongly follows a pattern: *Retrying* on most device errors and *Accepting* less when the system operates with high quality.

By observing the data of this interaction, we can also note that this user (Fig. 10) suspended the flow of conversation in many more cases – compared to other users – by being very selective, rejecting utterances very close conceptually.

3.3.2. Analysis of the Normal user type inference result

Fig. 11 shows one of the most challenging users to classify in our corpus. The system in this case takes over 24 turns to eliminate the Accommodating type, although it eliminated the Picky type from the 12th turn. Manual analysis of the data revealed that this user, despite being *Normal* in his average behavior, often exhibits *Accommodating* and sometimes *Picky* behaviors – crossing the boundaries of two types, thus causing the DBN to take longer to converge.

3.3.3. Analysis of successful user type inferences

In this subsection, we present the analysis of successful user type classifications suggested by the model (13 out of 15 interactions in our dataset were successful). Figs. 12 and 13b represent the identification of the

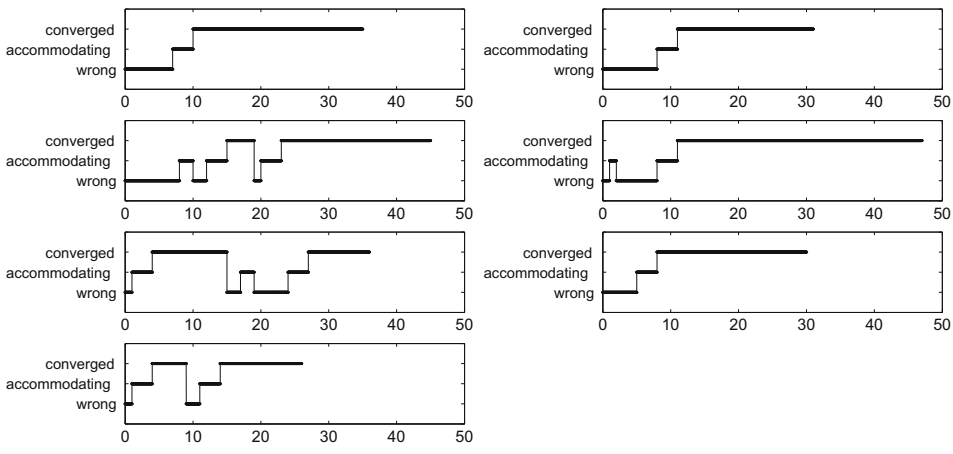

Fig. 12. Inference on the data of various "Accommodating" user types in the corpus. *X*-axis indicates the dialog interaction turn. *Y*-axis indicates three levels of prediction results – wrong, accommodating, and converged to Accommodating user types.

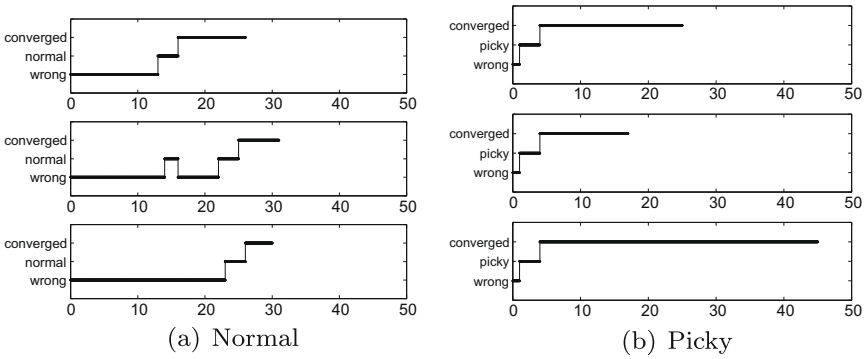

Fig. 13. Inference on the data of "Normal" and "Picky" user types over the dialog turns.

Accommodating and Picky user types. The correct user type is determined early in most cases (less than 10 interaction turns) even though some "Accommodating" users show different user types briefly in the middle of the whole interactions. The results imply that users in these two extreme types behave in their own style, especially when the system performance is low. And we can classify these two types early on by observing user behaviors and the system performance.

Different from the previous two extreme user types, the belief of "Normal" user type is gradually strengthened over turns by a tailing off of the other user types (Fig. 13a). This implies that it took comparatively more time to be at the midpoint, in terms of the number of retry/accept under low/high system performance, between the two extremes.

4. Online evaluation of user model

In the following sections, we report the results of online evaluation of the user model using agent feedback. For this purpose, our new speech-to-speech communication system (called *SpeechLinks*) was used, and the English speakers' user behaviors were analyzed. The design considered the following: Picky users tend to reject even small recognition errors which do not affect the overall meaning transfer from user-spoken utterance in the source language to machine-generated utterance in the target language. In the opposite situation, Accommodating type users tend to accept even critical recognition errors, which breaks natural conversations between users by causing incorrect meaning transfers through the device.

By providing agent feedback to users according to the user types, we could acquire better interaction efficiency (which will be defined in the result section) by encouraging users to change their behaviors in better direction.

4.1. Experimental setup

4.1.1. Participants and experimental domain

We recruited eight native speakers of English, four males and four females of ages between 20 and 28. All of them were undergraduate and graduate students at University of Southern California (USC). We also employed two Farsi speakers who are familiar with the *SpeechLinks* project. Farsi speakers were one male and one female, ages of 21 and 24 respectively, and also undergraduate students of USC. The choice of only two Farsi speakers familiar with SpeechLinks was made to reduce the variability space of the experiment.

In total, 32 interaction sessions were collected from eight native Speakers of English interacting with two native speakers of Farsi. For each interaction session, one native speaker of English and one native speaker of Farsi performed a diagnosis based on the provided scenario. The experimental time of each interaction session was approximately 30 min.

The domain of the experiment was medical diagnosis: Native speakers of English played the role of doctor and native Farsi speakers played the role of patient. Before the actual experiment, we gave a 1-h training

Diseases / Symptoms	Common cold	Flu	Food poisoning (Botulism)	Lactose intolerance	Depression	Insomnia
Abdominal pain	No	Mild, Upper left	Severe, Upper right	Severe, Upper left	Mild, Middle	No
Breathing	Normal	Difficult, Frequently	Difficult, Sometimes	Normal	Difficult, Sometimes	Normal
Chills	Slight, Frequent	Serious, Occasional	None	None	Slight, Occasional	Slight, Occasional
Concentration difficulty	Normal	Hard, Sometimes	Hard, Sometimes	Normal	Hard, Often	Hard, Often
Cough	Mild, Dry	Severe, Wet	No	No	Mild Dry	No
Diarrhea	No	Moderate, Sometimes	Intense, Frequent	Intense, Frequent	Moderate, Frequent	No
Dizziness	No	No	Severe, Irregular	No	Severe, Regular	Mild, Irregular
Exhaustion	No	Yes	No	No	Yes	Yes
Fatigue	Occasional	Often, Above the average	No	No	Often, Excessive	Occasional, Above the average

علایم عمومی
- تب بالا و سردرد

علایم مشخص (در صورت پرسیدن پزشک)
- خارش گلو
- احساس کوفتگی گاه به گاه
- خستگی خفیف، بعضی وقتها

علایم دیگر
- معمولی

General symptoms
- High fever and headache

Explicit symptoms (only if the doctor asks)
- Scratchy throat
- Occasional fatigue
- Mild tiredness sometimes

Other symptoms
- Normal

Fig. 14. Simplified example material: a part of doctor's diagnosis manual table for common cold (left). In the full size table, there are 12 diseases (column) and 30 symptoms (rows). A patient card for common cold presented to Farsi speakers (center), and its translation (right).

session to English speakers that included how to perform a diagnosis of the disease with the supplied materials: the doctor's diagnosis manual table (a simplified example is shown in Fig. 14 on the left) and the instruction of the experiment. The Farsi speakers were trained to use the system and to play the role of patient with the disease symptom card (simplified example in Fig. 14 on the center). The purpose of this experiment was to study the English speaker behaviors reacting to agent feedback (driven by the proposed model) rather than to study Farsi speaker behaviors. The goal of the English speakers (in the doctor's role) was to find out a disease of a patient in each interaction session. (The disease varies in each interaction session.) Four diseases (flu, SARS, depression and hypertension) were used equally for the eight English speakers during the experiment.

4.1.2. Scenario

The four scenarios were used in the same order during the experiment by each team (English–Farsi speaker pair). For each scenario, we provided a doctor's diagnosis manual table consisting of 12 diseases in the column and related symptoms in the rows. The diseases in the column were: common cold, flu, food poisoning, lactose intolerance, depression, insomnia, hypertension, high cholesterol, liver cancer, lung cancer, SARS, and diabetes. The symptoms in the rows were, for example: 'chills' and 'fatigue,' and the number of the symptoms was 30, in which the actual symptoms were varied depending on the disease. We built this table to be as realistic as possible using the medical diagnosis information from http://www.medicinenet.com.

Farsi speakers (patients) were given a symptom card which provided only a few symptoms of the disease. On the center image in Fig. 14, a symptom card for common cold is presented. We intentionally provided a few symptoms in each patient card to elicit more expressions from both speakers; English speakers needed to go through many combinations of diseases and symptoms in the look-up table to reasonably diagnosea disease on the symptom card of a Farsi speaker.

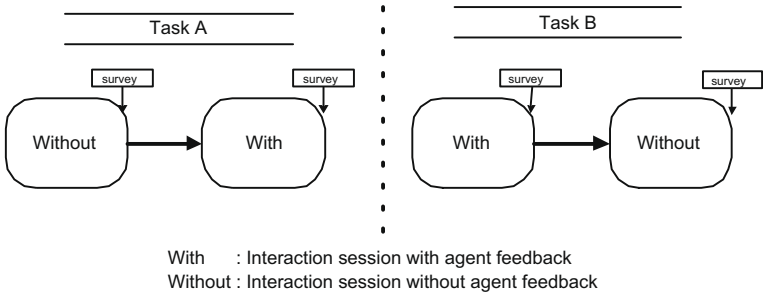

Fig. 15. All 8 English speakers performed both "Task A" and "Task B" with two Farsi speakers in different ways: four of English speakers performed "Task A" first and "Task B" later, and the other four performed in the reverse direction. Each English speaker met different Farsi speaker in the different Task.

Neither the English speaker in the doctor role or the Farsi speaker in the patient role knew the disease name during each interaction session. We informed them of the disease names at the end of all four interaction sessions.

4.1.3. Experimental procedure

The experiment was designed with two tasks, borrowing the idea of the evaluation method in the user modeling work by Komatani et al. (2003). Fig. 15 depicts this experimental procedure. In "Task A", native speakers of English performed the interaction session "without feedback" first and the session "with feedback" later. In "Task B", native speakers of English performed the interaction sessions in the reverse direction. In each task, the English speakers interacted with different Farsi speakers – one male speaker for one task and the other female speaker for the other task. For the tasks, each English speaker visited the experimental room twice (two days). We assigned the Farsi speakers evenly to the two tasks: each Farsi speaker participated in "Task A" 4 times, and the "Task B" four times. In total, we collected 32 interaction sessions from this experiment.

For evaluation purpose, we collected five different survey questionnaires from each participant during the experiment. One is the initial survey about demographic information of the participant and user perception on many topics (Appendix A), such as user type and error tolerance level and past speech interface experience. After each interaction session, a questionnaire was given to each participant for the evaluation of system performance along multiple dimensions (Appendix B), such as user satisfaction and interaction efficiency. In total, four evaluation questionnaires were collected from each participant. Detailed analysis of questionnaires is provided in Section 4.2.3.

Each session lasted for approximately 30 min — we gave a 5-min warning when the session was still continuing after 30 min. After finishing two sessions (with feedback and without feedback), participants gave us their opinions about the experiment.

All the interaction sessions were video taped. We analyzed the 32 interaction sessions in the video data in terms of identifying user types with their behaviors and user behavior changes and system performance.

4.1.4. Agent feedback for Accommodating and Picky user types

Two different wordings of agent feedback were prepared for two user types: Accommodating and Picky. When the system detected one of the two user types with high probability, it triggered the corresponding wording of agent feedback as in Table 7. The threshold of triggering an agent feedback was set as 0.65 which was acquired systematically from user training sessions. When the system detects either an Accommodating or Picky user type the first time, the wording (1) was presented to the users. After consecutive same user type identifications (e.g., three times), the system changed the wording, in this case, the wording (2) was presented to the users. The agent feedback was presented to users in this fashion throughout the whole interaction session.

User type identification was conducted by dynamic Bayesian reasoning as introduced in the Section 3.2. At each turn in the interaction, previous user behavior and ASR confidence level of the previous turn were utilized for computing the posterior probabilities of three user types. These probabilities were updated dynamically as the interaction proceeded.

Table 7
Actual wordings of agent feedback for two user types. Two different wordings were used alternately for the same user type in case of triggering the same agent feedback over and over.

	For Accommodating user type	For Picky user type
(1)	"Consider rejecting bad options and rephrasing"	"Accepting system errors, if those have little impact on meaning, may improve system performance"
(2)	"The system is not always right. Some errors can cause significant degradation in your communication. When presented with bad options consider rejecting them and retrying"	"The system often inserts some additional words in its recognition results. Consider accepting some errors if those affect little the concept of the recognized sentence"

The underlying assumption of the online experiment was that the ASR confidence level can be used to measure the ASR performance, which was measured offline by Word Error Rate (WER) as introduced in the section 3.2.1. The correlation between ASR confidence level and WER was mentioned and studied in Georgila et al. (2005) and Zechner and Waibel (2000). ASR confidence level was computed using features at multiple levels, such as weighted acoustic model and language model scores.

4.2. Experimental results

We present the results of our online experiment using subjective and objective measures from various sources: user interview, questionnaire, video analysis and log data analysis. Statistical analyses were performed with SPSS 15.0.

4.2.1. Subjective measure 1: user interview

The interview with participants gave us insightful information about user opinions about agent feedback and its relation to system performance. Participants told us that the agent feedback provided hints when the interactions went wrong and it helped for smoother conversation flows and information delivery. In particular, the participants commented that agent feedback helped in mitigating frustration caused by repetitive errors. One of the Picky type users said:

> "Agent feedback expedites conversation since users will not be repeating themselves in attempts to find an EXACT replication of their phrase."

4.2.2. Subjective measure 2: video analysis

By analyzing the video data of 32 interaction sessions, we subjectively identified user types of 8 English participants: seven participants were Picky and one was Accommodating. For this identification, we specifically investigated the behaviors of users when the machine-recognized utterances have functional words which do not affect on the whole meaning of the utterances.

The analysis of video data suggests a trend of user accommodation to system functionalities and errors. We observed that the participants became accustomed to agent feedback in the early turns of the interaction session, and in the later turns, they did not pay attention to the agent feedback. We conjecture that they already knew what the agent feedbacks were and when the agent feedbacks would be triggered. From this viewpoint, the users of "Task A" (interaction session from 'with feedback' to 'without feedback') seemed to cope with system errors better than the users of "Task B." More analysis in this regard is presented in the following section.

4.2.3. Subjective measure 3: questionnaire analysis

We collected five questionnaires from each participant, and the Likert-scale questions were given to the participants.

Table 8
The statistics collected from the Likert-scale questions of the initial survey given to the participants. We measured users' own perceptions about their ability to deal with general technology and speech interface, utterance length, and error tolerance levels.

Likert-scale questions	Mean	Std. dev.
Speech interface experience (0: none–10: more than 10 times)	5.94	4.23
Inclination for the general technology (0: never comfortable–10: comfortable)	6.81	1.51
Error tolerance level in the interactions with computers (0: not at all–10: completely)	4.88	1.96
Error tolerance level in the communications with humans (0: not at all–10: completely)	6.25	2.74
Utterance length (0: terse - 10: lengthy)	5.88	1.82
Hasty level when using computers (0: not at all–10: completely)	6.44	1.41
Ability to work with computers (0: worst–10: best)	5.63	1.31
Today's feeling (0: bad–10: good)	7.63	1.20

The initial questionnaire was intended to measure users' own perceptions about their ability to deal with general technology and speech interface, utterance length, and error tolerance level (Table 8). One finding from the initial questionnaire is that some users did not have speech interface experience at all, while others had some experience. To reduce this gap, we gave a one-hour training session to all the participants, which included how to use the system. Another interesting finding was that the error tolerance level of the participants was higher in the communication with humans than in that with computers.

Participants were given the evaluation questionnaire after each interaction session, and we collected four evaluation questionnaires from each participant who performed four interaction sessions in total. We measured user opinions in multiple levels, such as the system performance, user satisfaction and usefulness of agent feedback.

General user feeling (1: not at all–10: very much, standard deviation) about the interface of SpeechLinks indicates that the interface is intuitive (8.7(1.3)) and easy to learn (8.18(1.1)) but not foolproof (3.5(1.0)).

To measure the effect of agent feedback we present the comparison of user satisfaction (Table 9) between the interaction session with agent feedback, and the interaction session without agent feedback. This comparison was conducted between the two interaction sessions of each task (Task A and Task B) as described in Fig. 15. Higher user satisfaction was observed in the interaction session with agent feedback across the two tasks. More specifically, to find out statistical significance, a Paired Sample *T*-test was performed on each Task, and we acquired p values, 0.264 from "Task A", and 0.041 from "Task B". The observed significance level of "Task B" confirms the statistical difference between two interaction sessions ($p < 0.05$).

Basic statistics collected from the questionnaires which support the results of Table 9 are the following. Overall, user feeling about the usefulness (1: not at all–10: completely) of agent feedback was 6.5 (2.4) in "Task A" and 7.4 (1.7) in "Task B". The average number of triggered agent feedback per session was 7.1 (5.0) in "Task A" and 7.9 (3.6) in "Task B". The distraction levels (1: not at all–10: completely) of agent feedback in the two tasks (see Fig. 15) were 1.4 (1.3) and 1.7 (1.1) respectively. The topic difficulties (1: difficult–10: easy) in "Task A" and "Task B" were 5.7 (1.8) and 5.3 (1.4) respectively. User retry tendency (1: not at all–10: completely) in "Task A" was 6.8 (1.5) and that in "Task B" was 6.2 (2.1).

4.2.4. Objective measure: log data analysis

In this section, we investigated user behaviors accommodating to errors, and effects of agent feedback on the interaction efficiency.

Before presenting the results, it may be interesting to note some statistics collected from the two types of interaction sessions – with/without agent feedback. Averages (with standard deviation in the parenthesis) of session dialogue time were 33 min and 36 s (3 min and 2 s) with agent feedback, and 32 min and 27 s (4 min and 13 s) without agent feedback. Averages of the number of utterances in both sessions were 77.2 (26.6), and 70.0 (19.0), respectively. Averages of utterance length (in words) were 5.3 (1.5) and 4.6 (1.2), and averages of lasting time of each utterance (in seconds) were 4.2 (0.59), and 4.1 (0.37), respectively. Finally, overall number of agent feedback events in an interaction session was 10.7 (7.87) – excluding the interaction sessions without agent feedback.

In the video analysis results, we observed that on average only one participant was overall of the Accommodating type who endured relatively more recognition errors compared to the other seven participants. In the log data analysis, we investigated retry rates of the participants under low system performance, and seven

Table 9
Overall user satisfaction (Likert scale, 1: worst–10: best) after interaction session in each of the two tasks (with standard deviation). In "Task A", participants conducted an interaction session with agent feedback first and that without agent feedback later. In "Task B," participants conducted the interaction in the reverse order (without agent feedback first, with agent feedback later). Paired-Samples T Test shows that there is a significant difference in user satisfactions of two interaction sessions in "Task B" (5% level).

	Task A	Task B
First session	with: 7.0 (1.1)	without: 5.25 (1.7)
Second session	without: 6.0 (1.93)	with: 7.25 (1.3)
Statistical significance	$p = 0.264$	$p = 0.041$

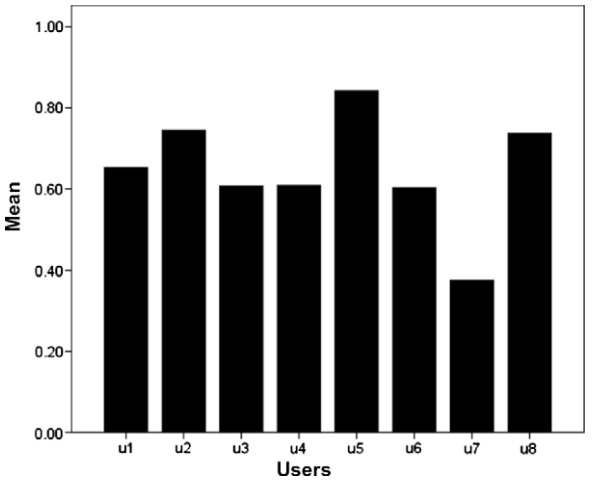

Fig. 16. User retry rates over the interaction sessions when the ASR performance is low. Interaction sessions without agent feedback were investigated. Seven users were observed as Picky and one as Accommodating.

users were observed as the Picky type, and one user was the Accommodating type (same as in the video analysis). The low quality (LQ) system performance is the region of low ASR confidence level. We investigated the interaction sessions without agent feedback for this analysis. The user retry rates over the interaction sessions are presented in Fig. 16.

One of the hypotheses in using agent feedback was whether we could increase the smoothness of the interaction. This interaction efficiency is highly correlated with the time the users behave in the Normal rather than Picky or Accommodating type regions. Normal type users are deemed not to be in the extreme to accept/reject system errors, so we expect to avoid extreme cases (such as severe repetitions or translation of large system errors) in their interactions. Intuitively, we assume smooth conversations when the participants are behaving more ''normal''. In our analysis of the data, the Normal user type was exhibited more during interaction sessions with agent feedback than during interaction sessions without agent feedback as shown in Table 10.

Another interesting aspect is to investigate the effect of agent feedback in improving user behaviors and in contributing to the efficiency of interactions. The agent feedback can be presented to users before the users catch the chain of same error situations. In this way, users can escape from the chain of possible error situations easily. Note that it is dependent on users to accept agent feedback and to use alternative strategies to recover from error situations. To illustrate the effect of agent feedback in this regard, we compared the percentages of user behavioral change from the previous turn during the interaction session without agent feedback and during the interaction session with agent feedback (Table 11). In this result, the user behavioral changes were counted only when the dynamic *Bayesian* reasoning identified two extreme user types (Picky and Accommodating) during the interaction session. In the interaction session without agent feedback, we triggered the agent feedback internally and observed whether user behavior changed from the previous turn or not. Note that there is a possible chain of errors when the two extreme user types are triggered by the dynamic *Bayesian* reasoning. As shown in Table 11, users changed their behaviors more with the help of agent feedback onscreen, indicating that the users had more chances to escape from a chain of error situations.

Table 10
Percentage (with standard deviation in parenthesis) of Normal user type that appeared during the two interaction sessions: with/without agent feedback. More Normal user type during the interaction sessions indicates more efficient interactions.

Without agent feedback	With agent feedback
0.37 (0.14)	0.44 (0.14)

Table 11
Percentages of user behavioral change from the previous turn under possible chain of errors during the interaction sessions without/with agent feedback. The changes of user behavior (accept/retry) were counted only when the dynamic *Bayesian* reasoning identified two extreme user types (Picky and Accommodating) during the interaction session. Note that two extreme user types were identified internally during the interaction session without agent feedback, and user behaviors were observed at this point.

Without agent feedback	With agent feedback
0.31 (0.21)	0.40 (0.16)

5. Discussion and conclusions

This paper addressed user behavior modeling approaches in a machine-mediated setting involving bidirectional speech translation. Specifically, usability data from doctor–patient dialogs involving a two-way English–Persian speech translation system was analyzed to understand two specific user behaviors. In addition to offline modeling results, data from an online experiment with agent feedback was performed and subjective and objective performance measures were reported.

We modeled user behavior with three user types, *Accommodating*, *Normal* and *Picky*. The granularity of user type can be adjusted according to the desired application. For example, classifying users in two categories, such as *Picky* and *Normal*, may work better when we do not want to take any steps for the case where the users are extremely tolerant of errors. In the offline data, we showed that one of the three types becomes obvious as a user maintains behavior consistent under the same conditions. This model can be utilized for the design of an efficient error-handling mechanism; in previous research (Prodanov and Drygajlo, 2005), a correct interpretation of the user's goal (intention) was helpful in dealing with errors in human robot dialogs. Ultimately, we believe that we can improve dialog efficiency and quality, task success, and user satisfaction that are important measures of success similar to past work on the PARADISE framework (Walker et al., 1997). In the online experiment, we addressed some of these issues with agent feedback being presented to users according to the model. High user satisfaction and interaction efficiency were reported in the interaction sessions with agent feedback.

There are several challenges that still need to be considered. One of the major challenges in empirically-based user modeling study is the availability of appropriate data. It is especially important to note that it requires a huge effort to collect, process and interpret the complex data from bilingual spoken interactions such as those considered in this study. It is well known that real human dialog data are complex to analyze and, due to the high degree of variance in the data, a large volume is required to create sufficiently accurate models. In terms of data size, more training data increase the accuracy of the test set (Tian et al., 2005). In addition, it is often unclear how much data is needed for optimal performance and what the appropriate features are to build a user model. These issues are of critical importance, especially when we attempt to model a user in a data-driven way.

In designing a mediated device, it is important to have a good understanding of the user model; thus to be able to appropriately modify the communication strategies, for example, by taking specific system initiatives. These system initiatives must be well founded on robust user models to ensure minimal user disruption. We designed triggering agent feedback in this fashion (to be not disruptive). However, some participants in the online experiment using agent feedback commented that they needed the feedback mostly in the early interaction sessions and repetitive feedbacks might turn out to be disruptive. How best to exploit the user model is still not a fully explored area, especially in light of partial observations (both temporally and qualitatively) of the user actions.

In the online experiment, we assumed that word error rate (WER) of the offline experiment can be substituted by ASR confidence level. This assumption is widely considered acceptable in the speech technology community. However, it is still debatable whether, under what conditions and with what features, we can accept this assumption.

We believe this work provides a first look and motivates future investigation of the benefits of user modeling in mediated, cross-lingual interactions. The advantages of this additional model in the system are becoming apparent, even at the infancy of speech to speech translation technologies. We believe that, as the devices mature, user awareness and mixed initiative will become even more critical.

Acknowledgements

This work was supported by the DARPA Babylon/CAST program, Contract N66001-02-C-6023 and by the DARPA TransTac program Contract Number NBCH1050027.

Appendix

A. Initial questionnaire

The questionnaire included the following questions from a scale of 0–10 as well as an "Identification number", "System number", "Age," and "Gender". The scale was graphical with the axis labeled and numbered, but for simplicity here we denote the two ends of the axis in brackets:

- A speech input interface is a computer program that recognizes your spoken words and responds to them. How many times have you used a speech input device before? ("None" to "More than 10/specify").
- How would you define your inclination for the (general) technology? ("Never comfortable" to "Comfortable").
- How much could you tolerate errors generated by machine while interacting with it? ("Not at all" to "Completely").
- How much could you tolerate spoken errors by human while communicating with him/her? ("Not at all" to "Completely").
- Some people are hasty to achieve a goal when use computers. Are you hasty in using computers? ("Not at all" to "Completely").
- How do you define your speaking style in terms of the length of sentence? ("Terse" to "Lengthy").
- Compared to other people I know my ability to work with computers is: ("Worse than most" to "better than most").
- How do you feel today? ("Bad" to "Good").

B. Evaluation questionnaire

The questionnaire included the following questions from a scale of 0–10 as well as an "Identification number", "System number", "Age," and "Gender". The scale was graphical with the axis labeled and numbered, but for simplicity here we denote the two ends of the axis in brackets:

- Overall, how satisfied were you in using the system? ("Worst" to "Best").
- How would you describe the interface of the system?("Difficult to use" to "Easy to use").
- The interface was: ("Rigid" to "Flexible").
- The topic of this interaction session was ("Difficult" to "Easy").
- Overall, did you find that the system perform well? ("Not at all" to "Very much").
- Was the interface easy to learn? ("Not at all" to "Very much").
- Was the interface intuitive to use? ("Not at all" to "Very much").
- How good was the system at delivering what you wanted to express? ("Not good at all" to "Extremely good").
- How often did you tend to retry? ("Not at all" to "all the time").
- Was the system foolproof? ("Not at all" to "Completely").
- The audio was: ("Choppy" to "Smooth").
- How many times have you seen feedback messages (please write or select one)? (choice of 0, 3, 6, 9, 12, 18, specify).
- Did feedback messages (if any) distract your interactions? ("Not at all" to "Completely").
- The time of showing a feedback message (if any) was ("Too long" to "Too short").
- Were feedback messages (if any) useful? ("Not at all" to "Completely").

- Did feedback messages (if any) help to expedite natural interactions? (“Not at all” to “Completely”).
- Did feedback messages (if any) help to induce accurate system performance? (“Not at all” to “Completely”).
- How do you think the system changed (Please answer if you have used this system more than once) (“Terrible” to “Wonderful”).
- Please describe how feedback messages were useful/unuseful.
- Do you have any suggestions for improving the system?

References

Batliner, A., Fischer, K., Huber, R., Spilker, J., Noth, E., 2003. How to find trouble in communication. Speech Communication 40 (1–2), 117–143.

Bauer, G.W., 1999. Interface for user/agent interaction. US Patent 5877759, March 2.

Black, A.W., Brown, R.D., Frederking, R., Singh, R., Moody, J., Steinbrecher, E., 2002. TONGUES: rapid development of a speech-to-speech translation system. In: Proceedings of the 2nd International Conference on Human Language Technology Research (HLT).

Carberry, S., Chu-Carroll, J., Elzer, S., 1999. Constructing and utilizing a model of user preferences in collaborative consultation dialogues. Computational Intelligence 15 (3), 185–217.

Conati, C., Gertner, A., Vanlehn, K., 2002. Using Bayesian networks to manage uncertainty in student modeling. User Modeling and User-Adapted Interaction 12 (4), 371–417.

Dash, M., Liu, H. Feature selection for classification. Intelligent Data Analysis 1 (3).

Eckert, W., Levin, E., Pieraccini, R., 1997. User modeling for spoken dialogue system evaluation. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).

Friedman, N., Geiger, D., Goldszmidt, M., 1997. Bayesian network classifiers. Machine Learning 29 (2–3), 131–163, Special issue on learning with probabilistic representations.

Georgila, K., Henderson, J., Lemon, O., 2005. Learning user simulations for information state update dialogue systems. In: Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech/Interspeech).

Grawemeyer, B., Cox, R., 2005. A Bayesian approach to modelling users' information display preferences. In: Proceedings of the 10th International Conference of User Modeling (UM).

Hall, J.A., Roter, D.L., Katz, N.R., 1988. Meta-analysis of correlates of provider behavior in medical encounters. Medical Care 26 (7), 657–675.

Horvitz, E., Breese, J., Heckerman, D., Hovel, D., Rommelse, K., 1998. The Lumiere project: Bayesian user modeling for inferring the goals and needs of software users. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence.

Jokinen, K., Kanto, K., 2004. User expertise modelling and adaptativity in a speech-based e-mail system. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL).

Kamm, C.A., Litman, D.J., Walker, M.A., 1998. From novice to expert: the effect of tutorials on user expertise with spoken dialog systems. In: Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP).

Kobsa, A., 1990. Modeling the user's conceptual knowledge in bgp-ms, a user modeling shell system. Computational Intelligence 6, 193–208.

Kobsa, A., Pohl, W., 1995. The user modeling shell system bgp-ms. User Modeling and User-Adapted Interaction 4 (2), 59–106.

Komatani, K., Ueno, S., Kawahara, T., Okuno, H.G., 2003. Flexible guidance generation using user model in spoken dialogue systems. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL), pp. 256–263.

Kuenzer, A., Schlick, C., Ohmann, F., Schmidt, L., Luczak, H., 2001. An empirical study of dynamic Bayesian networks for user modeling. In: Proceedings of the UM 2001 Workshop on Machine Learning for User Modeling.

Litman, D., Pan, S., 1999. Empirically evaluating an adaptable spoken dialog system. In: Proceedings of the 7th International Conference on User Modeling (UM).

Manavoglu, E., Pavlov, D., Giles, C.L., 2003. Probabilistic user behavior models. In: Proceedings of the 3rd IEEE International Conference on Data Mining.

Narayanan, S., Ananthakrishnan, S., Belvin, R., Ettaile, E., Ganjavi, S., Georgiou, P.G., Hein, C.M., Kadambe, S., Knight, K., Marcu, D., Neely, H.E., Srinivasamurthy, N., Traum, D., Wang, D., 2003. Transonics: a speech to speech system for English–Persian interactions. In: Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU).

Pakucs, B., 2003. Sesame: a framework for personalized and adaptive speech interfaces. In: Proceedings of the EACL-03 Workshop on Dialogue Systems: Interaction, Adaptation and Styles of Management.

Pitkow, J.E., Schuetze, H., Cass, T.A., Cooley, R., Turnbull, D., Edmonds, A., Adar, E., Breuel, T.M., 2002. Personalized Search, vol. 45, pp. 50–55.

Pitschke, K., 1994. User modeling for domains without explicit design theories. In: Proceedings of the 4th International Conference on User Modeling (UM), Hyannis, MA.

Precoda, K., Podesva, R.J., 2003. What will people say? Speech system design and language/cultural differences. In: Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU).

Prendinger, H., Mori, J., Ishizuka, M., 2005. Recognizing, modeling, and responding to users' affective states. In: Proceedings of the 10th International Conference of User Modeling (UM).

Prodanov, P., Drygajlo, A., 2005. Bayesian networks based multi-modality fusion for error handling in human robot dialogues under noisy conditions. Speech Communication 45 (3), 231–248.

Rich, E., 1979. User modeling via stereotypes. International Journal of Cognitive Science 3, 329–354.

Rich, E., 1999. Users are individuals: individualizing user models. International Journal of Human–Computer Studies 51 (2), 323–338.

Roter, D.L., Hall, J.A., 1989. Studies of doctor–patient interaction. Annual Review of Public Health 10, 163–180.

Sheinvald, J., Dom, B. Niblack, W., 1990. A modeling approach to feature selection. In: Proceedings of the 10th International Conference on Pattern Recognition.

Shin, J., Narayanan, S., Gerber, L., Kazemzadeh, A., Byrd, D., 2002. Analysis of user behavior under error conditions in spoken dialogs. In: Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP).

Tian, J., Nurminen, J., Kiss, I., 2005. Optimal subset selection from text databases. In: Proceedings of 30th Anniversary IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Tsiriga, V., Virvou, M., 2004. A framework for the initialization of student models in web-based intelligent tutoring systems. User Modeling and User-Adapted Interaction 14 (4), 289–316.

Walker, M., Litman, D., Kamm, C., Abella, A., 1997. PARADISE: a general framework for evaluating spoken dialogue agents. In: Proceedings of The Association for Computational Linguistics (ACL/EACL).

Yan, H., Selker, T., 2000. Context-aware office assistant. In: Proceedings of International Conference on Intelligent User Interfaces.

Zechner, K., Waibel, A., 2000. Minimizing word error rate in textual summaries of spoken language. In: Proceedings of NAACL-ANLP-2000.

Zhou, B., Gao, Y., Sorensen, J., Dechelotte, D., Picheny, M., 2003. A hand-held speech-to-speech translation system. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).

Zukerman, I., Albrech, D.W., 2001. Predictive statistical models for user modeling. User Modeling and User-Adapted Interaction 11 (1–2), 5–18.