USC-SIPI REPORT #391

User Modeling for Human-Machine Spoken Interaction and Mediation Systems

by

JongHo Shin

May 2008

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
Viterbi School of Engineering
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Suite 400
Los Angeles, CA 90089-2564 U.S.A.

USER MODELING FOR HUMAN-MACHINE SPOKEN INTERACTION AND MEDIATION SYSTEMS.

by

JongHo Shin

A Dissertation Presented to the
FACULTY OF THE GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(COMPUTER SCIENCE)

December 2007

Dedication

To my family.

Acknowledgements

I would like to thank thesis committee: Shri Narayanan, Ulrich Neumann, C. -C. Jay Kuo, and Panayiotis G. Georgiou. I would also like to thank Kevin Knight and Roger Zimmerman who were committee members for my qualification exam.

I would like to express my special thanks to all the USC Speech Analysis and Interpretation Laboratory (SAIL) people who helped me in data collection for my research.

My parents and brothers have been supportive all the time. Without them I would not have been able to finish my long trip to here.

Table of Contents

List of Tables

List of Figures

Abstract

One of the most fundamental challenges in building speech-enabled systems is "knowing the users." This information about users is captured in what is usually called a "user model." This study investigates user models for speech-enabled systems, which include both human-machine spoken interaction and machine-mediated human-human interaction systems. Because of the intrinsic error-prone property of statistical processing of human speech technology, errors are inevitable during the interactions to/through the speech-enabled systems. In this regard, this dissertation studies four different user models under uncertain error conditions of spoken dialog systems and spoken mediation systems. The user models were driven based on the data of mixed-initiative spoken dialogs, and multimodal (speech and visual) interactions of a spoken mediation system. The user models of this dissertation aims to contribute to accelerate the optimization of dialog management of the speech-enabled systems.

The addressed user models are about: (1) user behaviors under error conditions of a spoken dialog system; (2) multimodal user behaviors under uncertainty in two persons communication using a speech-to-speech translation system; (3) user behavioral changes over time in uncertain communication when using a multimodal interface of a speech-to-speech translation system; and (4) user level of tolerating errors implemented with a dynamic Bayesian network and possible *speech Accommodation* between two interlocutors. The model of dynamic Bayesian network was validated offline with the

multimodal interaction data of a speech-to-speech translation system, and online with *agent* feedback used in a multimodal interface of a speech-to-speech translation system.

ALEXANDER
No – wait – the moon's falling out of orbit – that's not possible!

VOX KIOSK (V.O.)
Well, considering it is, in fact, happening, I would assume it's
possible. The retrograde orbit began in 2005 when the demolitions
for the lunar colonies –

ALEXANDER
Why is it – breaking up?

VOX KIOSK (V.O.)
I was getting to that... The moon has reached the gravitational Roche
Limit, ... the nearest public evacuation shelters can be found at
Grand Central Station, Madison Square Garden.

People have dreamed, for a long time, of a machine which can recognize and react to humans using speech: where they freely converse with the machine, either a physical or a virtual entity. This includes situation in which people speaking different languages can communicate with one another without realizing the mediated translation machine. The machine recognizing humans would be able to act like a real human being, adapting itself to humans even in error conditions. It will not only communicate information but also understand the complicated contexts.

A human (Right) has a conversation with a virtual library hologram agent (left).
In the movie, *The Time Machine* (2002).

Chapter 1

Introduction

1.1 Significance of the Research

"Speech has long been a main tool for information exchange among human beings."

As speech technologies have evolved, people have expected speech-enabled systems to be increasingly intelligent and versatile. In particular, the development of a speech-enabled system (e.g., using a speech recognition interface) capable of effectively handling diverse user populations and languages has become desirable for many uses: it adapts to users effectively in multitude, such as user demographics (e.g., age or gender) or the level of proficiency when using the systems; it translates between different languages by considering each user's peculiar behaviors. The speech-enabled system equipped with the information about users provides enhanced user satisfaction as well as efficient user-adapted system functionalities.

The present study presents some of the important aspects of user models relevant for the two settings, human-machine spoken dialog interactions and human-human spoken interactions through a mediating device. In particular, the study limits the scope to four issues related to user behaviors in error conditions – users face various error situations when using the speech-enabled systems, and attempt to recover from those. Firstly, the policy of spoken dialog management needs information about users in error conditions. Users differ in the behavioral strategies to handle errors generated by the system, in particular, a spoken dialog system. Many dialog optimization approaches have been introduced without empirical validation of user models. Secondly, people emphasize meaning transfer when using the speech-enabled systems – how much information is transferred through the machine. This is a crucial challenge to address especially when users use a mediated speech translation system to interact each other in different languages. Thirdly, users change in their behaviors over time according to many factors, including system errors. These user changes are complex under the uncertain environment of speech-enabled systems. Fourthly, the speech-enabled systems need to be equipped with some strategies to deal with users differing in acceptance of system errors: some are more accommodating, and some not (that is, picky). This makes the systems more robust in terms of error handling. In the communication between interlocutors, the mutual effect on behaviors is another challenging topic to address.

1.2 Methodology

The information and assumptions about users are called "user models [44]." Inherently the user model is complex, and it is formed by combining multiple knowledge sources: user behaviors, psychological aspects of users, and purely mathematical assumptions about users. This dissertation investigates user models of the two speech-enabled systems by analyzing user behaviors profiled in the logs from the two systems. In particular, user models under error conditions of the two systems were empirically investigated. In the analysis, supportive user data (user opinion and observation on the user behaviors) were provided by survey questionnaire and the recorded multimedia data (video and audio).

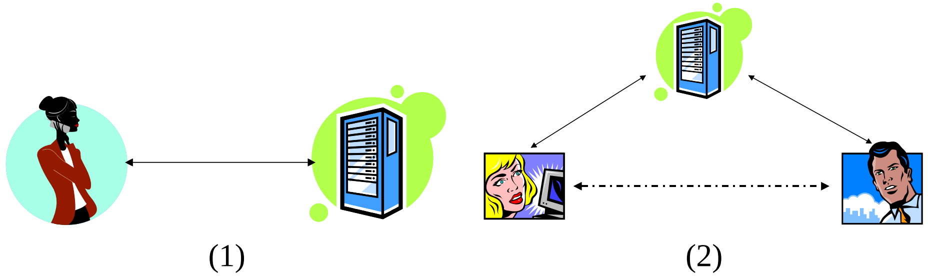
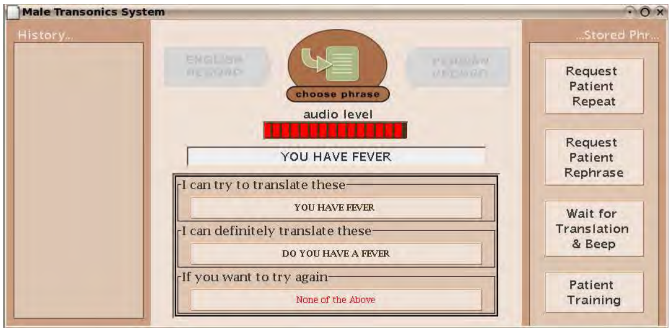


Figure 1.1: Two spoken interaction settings; (1) human-machine, (2) human-human with a mediating device. The first one has only one communication channel, and the second one has two communication channels – *the mediating channel* and *the interpersonal channel*.

Two speech-enabled systems used in this study are: a Spoken Dialog System (SDS) and a Spoken Mediation System (SMS), as illustrated in Figure 1.1. A Spoken Dialog System is a program supporting speech as an input modality to interact with a user. For example, the "Defense Advanced Research Projects Agency (DARPA) Communicator Travel Agency System [59]", which enables spoken communication between a human customer and a virtual travel agent program. The components of this system are: Automatic Speech Recognizer (ASR), Natural Language Processing (NLP), Dialog Management (DM) and Text-to-Speech (TTS). On the other hand, a Spoken Mediation System supports speech as an input modality for two interlocutors, and mediates conversations between them. One example system is the "Transonics Speech to Speech Translation System [64]", which enables spoken communication between two people who do not share the same language. The components of this system include all the components of SDS and Machine Translation (MT) additionally. In the SDS setting, the user and machine have conversations through one communication channel (direct human-machine channel), and in the other SMS setting, users interact through two communication channels (direct human-machine and direct human-human channels), as shown in (Figure 1.1).



YOU HAVE FEVER

I can try to translate these

YOU HAVE FEVER

I can definitely translate these

DO YOU HAVE A FEVER

If you want to try again

None of the Above

Figure 1.2: Example image of a multimodal interface of Transonics. After speaking, a user can choose one of machine-produced outputs presented onscreen.

The interface of speech-enabled systems incorporates not only speech input but also other modalities, such as visual. “Defense Advanced Research Projects Agency (DARPA) Communicator Travel Agency System [59]” used telephony speech interface (speech only), and “Transonics Speech to Speech Translation System [64]” used a multimodal interface equipped with speech and visual inputs. Figure 1.2 shows the multimodal interface incorporated in the Transonics system.

Speech-enabled systems are error prone. The errors arise from a number of sources: They can be speech recognition error, mismatched dialog strategy, mis-transferred meaning, or malfunctioning error. Even users may cause the errors by speaking off-topic. Example 1.1 shows a typical user-system conversation with error conditions in the DARPA Communicator corpus. In this example, the user gave up interacting with the system after two turns. This number of turns is statistically lower than the average number of turns with which users resolved the system errors in the DARPA Communicator corpus. We can call this user type a “quick-terminate.” This user type incorporates a strategy to finish the spoken interactions quickly in an error situation. Also the example 1.1 shows statistics of the user behaviors corresponding to the system responses, which include that the first user response was “repeat” when the system generated “sys-repeat” response in the previous interaction. When the majority of user behaviors is

(1) System said:	What is your destination
User said:	Phoenix Arizona
Recognizer heard:	ARIZONA [error, first time]
(2) System said:	What is your destination [*sysrepeat*]
User said:	Phoenix Arizona [*repeat*]
Recognizer heard:	ARIZONA [*error*]
(3) System said:	What is your destination [*sysrepeat*]
User said:	[uh] my god destination Phoenix Arizona [*frustrated*]
Recognizer heard:	++UH++ I'D GET DESTINATION ARIZONA LAST [*error*]
(4) System said:	What is your destination [*sysrepeat*]
User :	[*hang-up*]

Table 1.1: A portion of annotated dialog between a human and the machine from the DARPA Communicator corpus. The user first noticed that an error had occurred with the system cue, *sysrepeat*. User behaviors, such as *repeat*, *rephrase* and *frustrated* tend to co-occur with the system action *sysrepeat*.

"repeat" to the "sysrepeat", we could model user behaviors, which include an interaction pair of "repeat" and "sysrepeat." Through such approaches, we can build data-driven user models based on statistics acquired from interaction logs of speech-enabled systems.

User modeling starts with the analysis of user behaviors in the offline dataset, the logs of user interactions with a spoken dialog system (DARPA Communicator [59]). The study shows the analysis of user behaviors when things do not go well in the communication chain with the system. To investigate the situations, user behaviors under problematic conditions of real human-machine mixed initiative dialogs were annotated with tags identifying system error cues, recognizer errors and user behaviors. The first step in the analysis was to find out how users perceive the errors. This is important because the way in which errors are detected affects the amount of time it takes to get back on track. In addition, we investigated not only the statistics regarding successful user behaviors but also those regarding unsuccessful user behaviors. The analyzed user information is needed so that the system designers can determine strategies to overcome error situations effectively.

One of the major sources of error is an incorrect automatic conversion between speech and text. In such interactive applications, it is important that these errors do not impact the overall system performance in terms of meaning transfer between the user speech and the machine-produced output. In this regard, user behaviors were analyzed in terms of the number of concepts transferred through a mediating (translation) device. To clarify the definition of how many concepts are transferred in the utterances produced by the system (from original user utterances), the *Concept Matching Score (CMS)* was proposed. This score was defined based on "adequacy" levels, which assess manually the quality of translations by the system (proposed by Linguistic Data Consortium (LDC)). In addition, we compared machine translation performances between the unimodal (speech) interface setting and the multimodal (speech and visuals) interface setting by measuring the CMS. It is important for the system to be equipped with strategies to optimize the translation quality either through unimodal or multimodal user inputs.

Another important user modeling issue is user change. Previous user modeling studies assume that users get better over time in proficiency of using systems as they handle systems more. This user model is challenging to address given that users get affected by certain/uncertain factors over time, and the degree of that effect may change over time. In this regard, we empirically investigate user data collected over the four weeks of users' using a multimodal interface of a speech-to-speech translation system. The Concept Matching Score (CMS) was utilized in this study to measure meaning transfer from user speech to successful/unsuccessful machine-produced output. The study shows user behavioral changes over the weeks in terms of tolerating errors and user strategies dealing with errors. Supporting materials – user survey questionnaire, user interview, and recorded multimedia data (video and audio) – were used for additional analysis about users.

From the analysis of user behaviors under error conditions of Transonics above, it was observed that users differ in tolerating system errors: some users were more accommodating to the system errors than others. In this regard, we modeled user behaviors (retry/accept) with user types and speech recognition accuracy. The clustered users in the case studies were defined here in relative terms as *accommodating*, *normal* and *picky* user types. The user types and the features of the system – speech recognition accuracy and user behaviors (accept/retry) – were used to build a dynamic *Bayesian* network (DBN). This DBN keeps track of user types during the interaction turns. To validate the model, it was used to automatically determine user types in the offline experiment with the *10-fold cross validation*. To validate this model in real-time systems, online experiment was conducted with *agent* feedbacks. *Agent* feedbacks were provided for both *accommodating* and *picky* user types during the interaction turns, and interaction efficiency and user satisfaction were measured by the analysis of log data and user survey questionnaire. In addition, in the mediated device setting (as shown in Figure 1.1 (2)) of Transonics, we observed two channel communications; the mediated communication and the interpersonal communication. We attempted to address modeling interpersonal adaptation in *the interpersonal channel*, focusing on the design of a model based on the users' utterance length.

1.3 Contribution

The contribution of this dissertation is to address four aspects of user behaviors under uncertain error conditions of spoken dialog system and spoken mediation systems. The addressed user models are about: (1) user behaviors under error conditions of a spoken dialog system; (2) multimodal user behaviors under uncertainty in two person communication using a speech-to-speech translation system; (3) user behavioral changes over time in uncertain communication when users using a multimodal interface of a speech-to-speech translation system; and (4) user level of tolerating errors implemented with a dynamic Bayesian network and possible *speech Accommodation* between two interlocutors. The model of dynamic Bayesian network was validated offline with the multimodal interaction data of a speech-to-speech translation system, and online with the *agent* feedback used in a multimodal interface of a speech-to-speech translation system.

1.4 Limitations

In the dissertation, work outcomes are restricted to statistics acquired from data with human defined tags and transcription. This data-driven analysis and modeling in engineering has gathered increased attention recently because of its objectivity and practicality. Although we manually annotate and transcribe user data, it was conducted as objectively as possible. We conducted cross-validation among annotators with calibration sessions. Some part of the data was annotated with clustering techniques such as the *k-means* algorithm. This way, extendibility and objectivity were given high priority in the modeling process but we did not incorporate purely automatic annotation and transcription that may be used for other purposes.

A mediated device between two interlocutors brings about many issues which were attempted in this dissertation but not reported. For example, we investigated *speech accommodation* in prosody and acoustic levels between two interlocutors. However, it was not successful to find meaningful results. We believe the mediated machine between humans somehow prevents natural conversations. We leave these issues for future work to conduct research on how to improve that bottleneck.

1.5 Dissertation overview

The dissertation is organized as follows. Chapter 2 provides background and details of previous user modeling work. Chapter 3 reports user behaviors under problematic conditions of spoken dialog. In Chapter 4, analyzed multimodal user behaviors is presented, which use a measure of meaning transfer of a speech-to-speech translation system. In Chapter 5, user behavioral changes over several weeks are analyzed in the data of users using the multimodal interface of a speech translation system. In Chapter 6, the user type model in regards to the speech recognition error tolerance levels is presented. A dynamic Bayesian network is used as an inference mechanism of user types over time, which is evaluated with both offline and online experiments.

Chapter 2

Background and Previous Work

2.1 Spoken Interactions in Two Settings

In this section, two cases of spoken interaction system architecture are introduced: Human-Machine and Human-Human with a mediating device. These provide the proposed experimental environment for analyzing and modeling users of such systems.

2.1.1 Spoken Dialog System

A Spoken Dialog System (SDS) is a program supporting speech as an input/output modality to interact with a user. Speech-enabled Human-Machine interaction applications have matured for decades. Those applications evolve and cover a wide range of fields. For example, a "travel agency system" [59] provides itinerary services to a user, a "help desk" [19] application plays a receptionist role in routing a call to human agents or departments, and an "in-car speech recognition system" [96] offers a flexible and comfortable driving environment for a driver. As the typical components in SDS, the modules of the "Communicator" system [59] are depicted in Figure 2.1 with description below.

- Automatic Speech Recognizer (ASR): receives a digitally encoded acoustic signal and generates a string of words as output.
- Spoken Language Understanding (SLU): gets a word graph and generates the most probable word sequence and meaning as output.

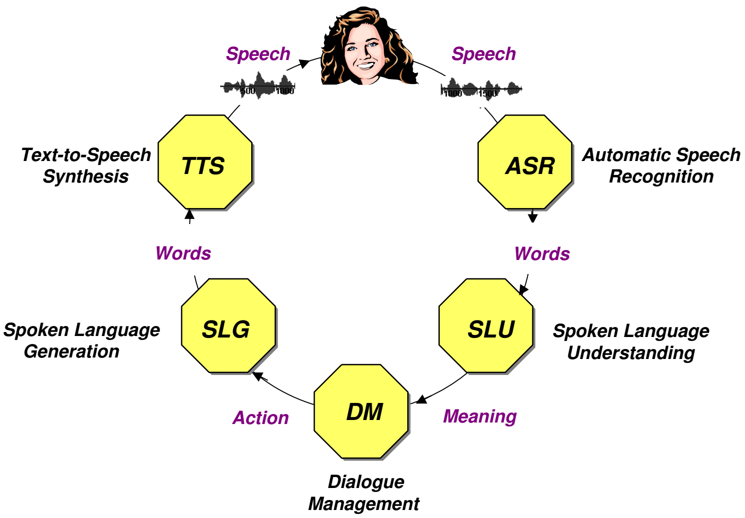


Figure 2.1: An architecture for Spoken Dialog System (SDS).

- Dialog Management (DM): monitors dialog flow between human and machine, collects information from a user, and interacts with database and decides where information is needed most.
- Spoken Language Generation (SLG): produces natural language text from computer-internal representations of information.
- Text-to-Speech (TTS): converts text into audio output.

2.1.2 Speech Mediation System

A Speech Mediation System (SMS) is a program designed to mediate natural speech interaction between two interlocutors, particularly when equipped with "speech translation capability". Recent trends show that advanced research institutions have been developing the systems under this setting. In our group, "Transonics" is such a system, which mediates a medical domain conversation between an English speaking doctor and a Persian speaking patient [64]. A portable speech-to-speech translation system was developed between English and Croatian speakers [7, 105]. "VerbMobile" [98] was developed for the domains of appointment negotiation, travel planning and hotel reservation.

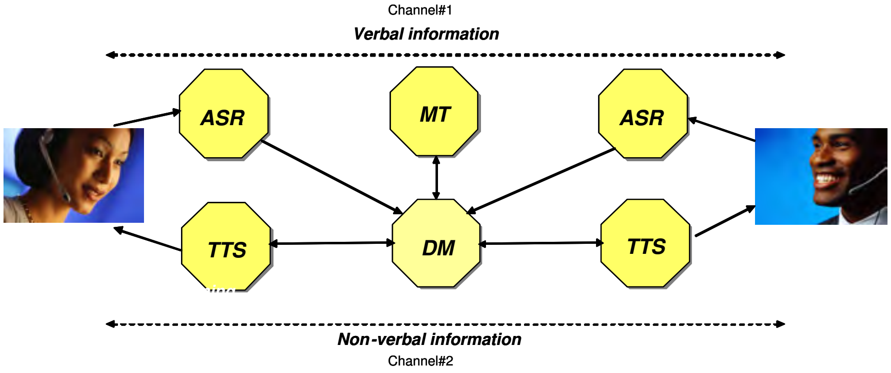


Figure 2.2: An architecture for a translation driven Spoken Mediation System (SMS).

It processes multilingual spontaneous speech (English, German and Japanese). Even though detailed implementations are different, most of developed systems include typical components such as ASR, MT, DM, and TTS. The typical architecture is depicted in Figure 2.2.

Most systems include:

- Automatic Speech Recognition (ASR): processes user speech signal with a target language (English/Farsi) and generates words of corresponding language as output.
- Machine Translation (MT): translates text between natural languages (e.g.English and Farsi).
- Dialog Management (DM): conveys translated spoken words to the output components such as TTS/GUI. Also, this monitors information flows among users and the device.
- Text-to-Speech (TTS): converts the translated text into spoken voice output.

Spoken interactions of SMS, compared to those of SDS, can be defined as two channel communication. As depicted in Figure 2.2, verbal information goes through one channel and non-verbal information such as a gesture, a eye-gaze, or a cultural adaptation of two users can be extracted from the other channel.

2.2 Previous User Modeling Approaches

"User Modeling" is an interdisciplinary research topic aiming to define the information about a user. This fundamental idea of it is applicable and reusable in other applications. In this chapter, the key contributions of relevant previous user modeling research are summarized. We first introduce general user modeling work. After this, specific user modeling research in the setting of Spoken Dialog System and Spoken Mediation System are outlined.

2.2.1 A General User Model

"Stereotype" user model is known to be the most typical one in the user modeling community. Basically, it is a collection of frequently occurring characteristics of users and covers a wide range of its applications. Rich [84] first proposed this mechanism for diverse individuals based on distinctive or different personalities, goals, and so forth. With the stereotype user model, the system can react to different users effectively with little information about them. Although initial values are hand-crafted and the feature selections are heuristic, it shows a productive user model in general as a pioneer work.

Kobsa introduced a generic (which means "domain" or "application" independent) user modeling shell system [43, 46]. It allows assumptions about the user or stereotypical user groups to be represented in a first-order predicate logic and assists the adaptation to the current users by taking the user's presumed knowledge, beliefs, and goals into account. The shell developers decide the structures and processes based on the intuition and experience from prior work on user-adaptive systems. Kobsa defined the requirements of representational and inferential expressiveness that can be a summary of various user modeling approaches [45]. These are:

- Representation of assumptions about one or more types of user characteristics in models of individual users (e.g. assumptions about their knowledge, misconceptions, goals, plans, preferences, tasks, and abilities).
- Representation of relevant common characteristics of users pertaining to specific user subgroups of the application system (the so-called stereotypes).
- Classification of users that belong to one or more of these subgroups, and the integration of the typical characteristics of these subgroups into the current individual user model.
- Recording of users' behavior, particularly their past interaction with the system.
- Formation of assumptions about the user based on the interaction history.
- Generalization of the interaction histories of many users into stereotypes.
- Drawing of additional assumptions about the current user based on initial ones.
- Consistency maintenance in the user model.
- Provision of the current assumptions about the user, as well as justifications of these assumptions.
- Evaluation of the entries in the current user model, and comparison with given standards.

Predictive statistical user models are getting attention these days [108, 109] and these cover the various aspects of human behavior, such as goals, actions, and preferences. The main evaluation metrics of these models are recall and precision, predicted probability and accuracy, and utility. The specific models include *linear*, *Term Frequency Inverse Document Frequency (TFIDF)*, *Markov model*, *Neural network*, *Classification methods*, *Rule induction*, and *Bayesian network*.

2.2.2 User Models in the Human-Machine Spoken Interaction Setting

- *Spoken Hyperarticulate and Disfluency*

Coping with user's spoken *Hyperarticulation* and *disfluency* help the system in dealing with various kinds of human speech. Oviatt studied hyperarticulating [73] and disfluencies [65] in human speech under a Human-Machine spoken interaction setting. The basic assumption is that a human speaks differently in the Human-Machine interaction setting compared to Human-Human interactions; A human (1) hyperarticulates more as the speech recognition error increases, (2) has fewer disfluencies with short and structured utterances.

Oviatt suggested a model of "hyperarticulate" with the two-stage CHAM (Computer-elicited Hyperarticulate Adaptation Model). With a high error rate in the experiment, the human speech includes more hyper-clear phonological features, fewer disfluencies and changes in fundamental frequency. Based on the observations under error situations, CHAM model specifies that users' speech will adapt to the linguistically-specified hyperarticulation profile as in Table 2.1.

Pause interjection	+92%
Pause elongation	+75%
Disfluencies	-53%
Intonation final fall	+19%
Speech elongation	+12%
Hyper-clear phonology	+9%
Pitch minimum	-2%
Pitch average	-1%

Table 2.1: Summary of relative change in linguistic dimensions of hyperarticulation [73]. All magnitudes shown represent statistically significant change during repetition.

For a “disfluency” model, Oviatt suggested a simple linear regression model. Higher disfluency rates were related to the length of utterance and lack of structure in the presentation format. When users’ utterances become lengthier, there occurs the higher possibility of disfluency. Also, it turned out that structural speech lead fewer disfluencies.

Regarding the disfluency, Shriberg [95] investigated and modeled it with *filled pauses, repetitions, repairs* and *false starts*. The model of disfluencies only utilized the prosodic features – *duration, distance from pause* and *f0*. All the features are related to the disfluency model with the *decision tree*.

* *User Goal or Intention*

A user goal or intention has been regarded as one of the most important user aspects in user modeling. It gives the system an assumption of a user’s limited behavior patterns based on a goal or intention. Therefore, it allows the system to reduce its burden of determining all the possibilities in user behaviors. Allen [4] pioneered this area by analyzing user utterances. It is claimed that a user expresses his or her goal information in the utterances. Recently, Horvitz put a lot of effort into this topic under uncertain conditions; harnessing models of users’ goals and needs [32, 30]. By utility-directed procedures, probabilistic relationships among intentions and spoken utterances are modeled and user actions are inferred (with the largest expected utility) through a Bayesian network.

* *User and Machine Initiative Modeling*

Spoken dialog systems adopt the “initiative” model to give the best freedom of interruptions to a user in spoken interactions. Also, it allows the system to interrupt the interactions whenever needed. Allen [3] seems to be the first researcher to deal with this issue with a taxonomy of mixed-initiatives. He proposed specific mechanisms such as contextual interpretation, turn taking and grounding. Horvitz [31] approached this issue with a more complicated condition, “under uncertainty”. Both hand-built and automatically-learned probabilistic user models (Bayesian networks) give the system the ability to take actions efficiently (mixed-initiative). Chu-Caroll [10] distinguished between *task* and *dialog* initiatives and presented a model for tracking shifts in both types of initiatives in dialog interactions. Chu-Caroll showed how to track the shift or lack of shift in task/dialog initiatives by the eight cues: *Explicit requests, End silence, No new info, Questions, Obligation fulfilled, Invalidity, Suboptimality, Ambiguity*. She utilized the Dempster-Shafer Theory to model tracking initiative between a human and a machine. New task/dialog initiative indices between a human and a machine were computed based on the current indices. The current observed cues were utilized with the current indices to determine the next task/dialog initiative holders.

* *User Modeling Framework*

User modeling middle-ware systems are effective in rapidly building an application which can handle diverse users. There have been some general frameworks in HCI community such as BGP-MS by Kobsa [46], GUMS system by Finin [21], UM by Kay [41]. These are general in their utility by allowing a wide range of representations for an individual user. Particularly Pakucs presented a framework for speech interfaces [74, 75] which adapt to a user by recognizing a context such as *location*, *time*, or *activity*. The system architecture is generic and task-oriented and it utilizes a feature vector of contexts.

* *Applications with a User Model*

One of the successful spoken dialog applications with a user model was built by Komatani [47]. Its domain is a “Kyoto city bus information kiosk.” Implemented user models include *skill level to the system, knowledge level on the target domain* and *degree of hastiness*. A decision tree with features obtained from a single utterance and dialog sessions is used for a user modeling. Annotated tags are hand-crafted and the values of a model are learned from the system's real dialogs.

Another important application requiring a user model is a tutoring system. It needs to know what knowledge the student has and what goals the student wishes to achieve. Conati [12] implemented "Andes" which provides long-term knowledge assessment, plan recognition and prediction of student actions. The probabilistic modeling technique (a Bayesian network) was used in the modeling of "Andes." The uncertainties defined in the design of "Andes" include *Context specificity, Guessing, Mutually exclusive strategies, Old evidence, Errors, Hints, Reading latency, Self-explaining ahead, Self-explanation menu selections*. As an evaluation metric, Conati showed the overall system performance enhancement by adopting the "Andes" to the tutoring system.

2.2.3 User Models in the Human-Human Spoken Interactions with a Mediating Device Setting.

User modeling work for Human-Human interactions with a speech-enabled mediating device is a relatively unexplored area compared to that for the Human-Machine spoken interactions. Some speech-to-speech translation applications such as Verbmobil [98] mediated the speech between two speakers who are using different languages but no user modeling research was done in the development of the system.

Computer Mediated Communication (CMC) community has conducted some research for user modeling in the clear communication between people under a mediating device setting. In this regard, the clear communication means no noisy operations such as speech recognition and machine translation (which are basically error-prone).

As one of the significant researchers, Isbister explored the cultural issues between Americans and Japanese with a helper agent in virtual meeting space [34]. The research issues include *perception of each other, each others' national group* and *the effects on their style of behavior*. This kind of work inspires us to investigate the social relationship between people in spoken interactions under a mediating device setting, which is a noisy mediating environment for the spoken interactions.

Recently, Shriberg [94] described four fundamental properties of spontaneous speech when speech recognition technology is involved: *punctuation, disfluencies, turn-taking and user state (or emotion)*. These four phenomena suggested multiple levels of speakers' information contained in behaviors and speech itself (lexically and acoustically). Shriberg proposed that we need to deal with these four phenomena to improve the performance and the utility of intelligent spoken language applications, as well as increased scientific understanding of natural speaking behavior.

On the other hand, in Cognitive Science, Linguistics and Phycology communities, it is claimed that natural Human-Human spoken interactions include the information of *"Accommodation"* between interlocutors. It can be either verbal such as *words and prosody*, or non-verbal such as *gesture and emotion* accommodation. Participants of the interactions concern about mutually combined social standing, task goals and communicational efficiency. This *Accommodation Theory* can be utilized to improve a mediating device performance and user satisfaction by encouraging users to interact naturally as if there were no machine involved.

Theoretical supports for the accommodation between interlocutors in the *interpersonal channel* are based on the *Speech Accommodation Theory* (Giles [26]). Giles divided the *Accommodation* into two strategies : *convergence* and *divergence*. The former refers to the processes whereby two or more individuals alter or shift their speech to resemble each other's speech. The latter indicates the ways in which speakers accentuate their verbal and non-verbal differences in order to distinguish themselves from others. Both are deployed by individuals to signal identification with, or dissociation from, the communication patterns of others. The Ultimate goal by adopting this *Speech Accommodation Theory* is to ensure successful communication. This achievement can be defined by increased level of satisfaction and conversation efficiency. This potential achievement inspires the application of *Speech Accommodation* in various speech-mediated device designs. In a medical domain, for instance, the satisfaction and the efficiency are critical metrics in measuring performance of communication between a doctor and a patient [87].

Fais [20] reported lexical accommodation studies in machine-mediated spoken interactions and incorporated three different settings: Human-Human monolingual, Human-Interpreted bilingual [1] and Machine-Interpreted bilingual. Significant accommodation results were found in all three settings that support the fundamental theories described in Chapter 6. The underlying hypotheses suggested in this context are:

- In human-human interaction, we should find significant lexical accommodation.
- The human-interpreted setting constitutes both a Human-Human interaction and a more stressful communication environment, one in which communicational efficiency is a concern.
- The machine-interpreted setting only indirectly involves Human-Human interaction; all dialog is mediated by the "machine" interpreter.
- We expect that clients will accommodate the machine to some extent, that clients' word choice will be affected by their perception of "what works," or "what the machine knows."

[1] English-Japanese

2.2.4 User Modeling on Multiple Modality Usages

Communication or information exchanges by using multiple modalities such as pen, speech or keyboard is referred to as *multimodal communication*. The research on user modality usages of multiple modalities expedites the design of an efficient speech-enabled interface. The effective user model on multimoality supports users in managing cognitive load [69]. Oviatt [66, 68, 72] investigated this issue extensively in terms of

- User preference
- The situation when users interact multimodally
- Integration and synchronization patterns
- Individual differences
- Complementarity and redundancy
- Performance and linguistic efficiency
- Error suppression

The assumption is that people can interact freely, differently and multimodally (e.g., speech, gaze, gesture, pen, visual, etc). Oviatt claimed that speech interface only is not enough for the robust interactions between the system and a human; multimodal approach expedites error handling and reduces a communication barrier such as hyper-articulate.

2.2.5 User Simulation for Spoken Dialog System (SDS)

Along with the stochastic approach to develop a Spoken Dialog System, the user simulation approach has been researched to improve the system performance in terms of less expensive task, less labor and fewer errors. The system is designed to generate most efficient responses to a user based on the simulation results.

User Model	Informal description
Reference	a first try to construct a “reasonable” behavior; all probabilities selected according to common sense.
Patient	a reference user with nearly infinite patience that would hang up the phone after 99 turns; all dialogues are expected to lead to success.
Submissive	questions will be answered, but no additional information will be volunteered.
Experienced	much more over-informative than the reference user; gives information on her own, with slightly higher patience.

Table 2.2: Characteristics of experimental user models (Eckert [17]).

* *A Conditional Probabilistic User Model*

Eckert [17] first introduced this user simulation concept in speech community to assess the dialog system performance. The user response (e.g., intention) is modeled with a system stimulus by the conditional probability,

$$p_{i,j} = p(U_t = I_i | S_{t-1} = I_j) \tag{2.1}$$

where, I_i and I_j denotes sets or sequences of intentions. S_{t-1} is a system stimulus at time $t-1$ and U_t is a user response at time t. Also, user initiative is modeled by the probability $p(U_t = I_i — S_{t-1} = \varepsilon)$ of presenting I_i in response to an open ended question.

Based on this simple probabilistic user model, Eckert [17] tested the system with handcrafted (partially learned from the corpus) values for each user type as described in Table 2.2.

* *A Constrained Probabilistic User Model*

In the domain of air travel information systems, Levin [54] simulated a user who learns dialog strategies automatically in an efficient way. Reinforcement learning was

Constrained User Model	Description
Response greetings	Probability $P(n)$, n = 0,1,2,..., $P(attribute)$, where $attribute$ = ORIGIN, DESTINATION,..., and the probability distribution on the value of each attribute [e.g., $P(Boston|ORIGIN)$, $P(Delta|AIRLINE)$].
Response to constraining questions	$P(k_R|k_G)$, i.e., the probability of the user specifying a value for attribute k_R when asked for the value of attribute k_G.
Response to relaxation prompts	$P(yes|k_G)$ = 1 - $P(no|k_G)$, i.e., the probability of accepting (or rejecting) the proposed relaxation of attribute k_G.

Table 2.3: The simulated user model characteristics in Levin [54].

used to estimate the optimal dialog strategy and supervised learning was used to estimate a user model. Levin assumed that a user response depends only on the current system action and it is constrained by some attributes. The simulated user is parameterized as in Table2.3. The assumption is that there is no recognition and understanding errors.

* *A Probabilistic User Model with Goal Information*

By adding "Goal" information to the Eckert's pure probabilistic user model (this concept was initially introduced by Scheffler and Young [90] to keep goal consistency), Pietquin [77] explicitly model the dependencies between a user's actions and his goal:

$$P(provideA_t|requestA_t, goal) \qquad (2.2)$$

where, A_t = user actions at time t.

User goal was represented by a simple table of attribute-value pairs and the probabilities are handcrafted. This model was not evaluated with the real data or the system.

* *User Simulation with a Linear Feature Combination and N-gram Language Models*

Georgila [25] recently introduced a user model based on the dialog information states with linear feature combination (Equation 2.3) and n-gram [2]. In this context, n-gram treats a dialogue as a sequence of pairs of speech acts and tasks. The performance of user models with linear combination and n-gram performances are measured by comparing the perplexity [3] and evaluated against the system policy.

$$P(a|s) = \frac{exp(f(s)^T w_a)}{\sum_a exp(f(a)^T w_a)} \tag{2.3}$$

where, w_a = weights being trained on the state-action pairs from the training data.

$f(s)$ = vector of real valued features for the state s. a= a user action.

* *Evaluation Metrics of User Simulations*

In [100], user satisfaction and system performance was measured with the PARADISE evaluation framework. Recently, Schatzmann [88] quantitatively evaluated three user models introduced above: bigram user model (Eckert [17]), user models with constrained attributes and the goal information (Levin [54] and Pietquin [77]). This assesses how realistic the best response is that the simulated user can predict. The precision and recall rates of the three simulated user models are measured and the goal completion rates are compared with those of the pre-existing systems.

[2] Originally n-gram is a subsequence of n letters from a given string after removing all spaces from language modeling community. For example, the 3-grams that can be generated from “good morning” are “goo”, “ood”, “odm”, “dmo”, “mor” and so forth.

[3] the state of being perplexed.

Chapter 3

Analysis of User Behavior under Error Conditions in Spoken Dialogs

"User behaviors contain abundant information about users."

We focus on developing an account of user behavior under error conditions, working with annotated data from real human-machine mixed initiative dialogs. In particular, we examine categories of error perception, user behavior under error, effect of user strategies on error recovery, and the role of user initiative in error situations. A conditional probability model smoothed by weighted ASR error rate is proposed. Results show that users discovering errors through implicit confirmations are less likely to get back on track (or succeed) and take a longer time in doing so than other forms of error discovery such as system reject and reprompts. Further successful user error-recovery strategies included more rephrasing, less contradicting, and a tendency to terminate error episodes (cancel and startover) than to attempt at repairing a chain of errors.

3.1 Introduction

Modeling human-machine spoken dialog interactions is gaining a lot of attention [54, 59] with the recent deployment of several complex dialog systems, for e.g., [101, 53, 106]. An important aspect of this problem is the understanding and modeling user behavior to enable realistic optimization of dialog strategies. It is well known that many of the underlying components of the state-of-the-art dialog systems such as automatic speech recognition and understanding rely on data-driven statistical models and, in general, are prone to errors of varying types and extent. In addition, there are other possible systems and user induced errors. Our work targets user behavior modeling under such error conditions in the context of human-machine spoken dialogs.

The DARPA Communicator spoken dialog systems, implemented at several sites, represent some of the most recent advances in the design of mixed-initiative spoken language systems [101, 53, 59]. The availability of transcripts of realistic spoken dialogs from some of those systems provides an excellent opportunity to investigate the behavior of human and machine interactions in mixed-initiative dialogs. In the present work we set out to understand the dynamics of user behavior under system errors and how the combination of system errors and user reactions to them affect the ultimate success of a dialog. In preparation for this study, we annotated a portion of the June 2000 Communicator dialogs for several features, including a categorization of both user and system behavior. The data and the extended annotation scheme are described in section 2. The results of our study are described in section 3. The paper concludes with a summary and discussion of the results in section 4.

3.2 Data and Annotation

The data used were the orthographically-transcribed travel arrangement dialogs from the DARPA Communicator project recorded in June 2000. Each dialog consists of some number of exchanges between a computer travel agent and a human and is represented as a three-line triple consisting of a system utterance, a user utterance (manually transcribed from recordings), and what the ASR system heard and provided as input to the dialog system. The data and the collection procedure are described in detail in [59]. In the Communicator dialogs, 85 experimental subjects interacted with 9 different "travel agent" systems. Out of the 765 possible dialogs, many are empty, or contain no user participation. We worked with about 141 of those total dialogs (that consisted of at least 1 turn). The average length of these dialogs was 18 exchanges. The amount of data is comparable to the data considered in a similar study by Aberdeen et al [1].

3.2.1 Tagging

Following a review of the recent work on analysis of human computer dialogs, we devised a tagging scheme consisting of 23 tags with which to monitor 3 dimensions of the dialogs: user behavior, system behavior, and task status. Since our goal was to do a quantitative analysis of the (disruptive) effect of errors, existing tagging schemes, while instructive, were not directly applicable. Automatic analysis of error conditions beyond the ASR word error rate is difficult without the aid of manual tagging. Hence, manual tagging was necessary. However, for example, unlike [1], we do not keep track of the subtask in which the error occurred, nor do we distinguish between dialog acts as in citeWalkerDialogTags. Finally, the user utterances in the communicator data are very short, averaging 3 words. Under these circumstances, we also have not made an to attempt labeling disfluencies as projects dealing with longer, more open-ended utterances have done [56, 2, 51]. The detailed tag set together with usage conventions and examples of application are provided in Appendix A. Briefly, the tag set for our purposes included (1) SYSTEM tags: explicit confirmation, implicit confirmation, help, system repeat, reject, non sequitur (2) USER tags: repeat, rephrase, contradict, frustrated, change request, startover, scratch clarify, acquiesce, hang-up (3) TASK tags: error (at the recognized utterance), back on track, task success. For error segments, we locate the beginnings of errors, and place a generic "error" tag on the ASR output that resulted in an error (Note that the standard ASR word error rate for each turn is also calculated). Within error segments we focus on three phenomena: system utterances which exhibit a system reaction to the error, user utterances which react to or try to correct the error, and the means by which the user becomes aware of the error. Sometimes the user becomes aware of an error because of a system rejection such as, “I’m sorry, I couldn’t understand you.” or a verbatim repetition of a system prompt for information. Other times implicit confirmations or non sequiturs in system utterances alert the user to the presence of an error, in which case the user must try to make the system aware of the error. Because the scenarios were conducted by paid subjects arranging for hypothetical travel for this particular data collection, some users had a tendency to acquiesce to errors that proved difficult to correct, or even to change the nature of the travel request in response to repeated recognition errors. These deviations from the original plan are also marked. Finally, we tag the point at which the dialog gets back-on-track (BOT), marking the system utterance in which the user could reasonably discover that the portion of the task derailed by an error has been successfully understood. At the end of the dialog we indicate whether the arrangements were successfully completed or ended in a hang-up or acquiescence to some error. The tagging was done by two annotators and showed 87% inter-annotator agreement. The tagging conventions used allow the assignment of all applicable tags to the dialogs. The agreement measure used was the number of identically tagged lines, divided by the number of lines reviewed and tagged. The measure is conservative in that it counts as agreement cases where 100% identical tagging appears on exactly the same line for both annotators. It does not include partial overlap, or positional offset. Following the tagging itself, we analyzed the dialogs and user histories from several perspectives, seeking patterns in user behavior, and correlations between user behavior and the length and severity of error segments.

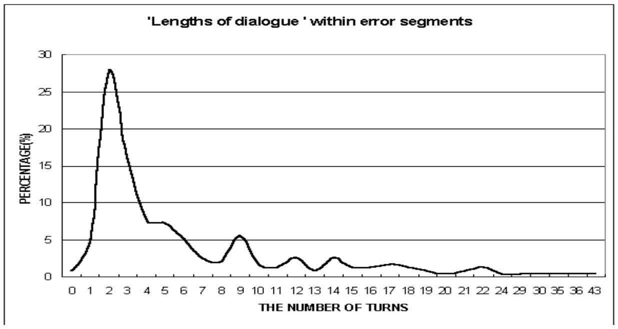


Figure 3.1: Normalized histogram of the length of error segments (number of turns).

3.3 Results and Discussion

Firstly it is useful to get a general sense of the presence of errors in the dialogs. The data, overall, is dominated by errors of various types. The roughly 2528 turns we tagged consists of 141 dialogs conducted with 35 paid subjects. The dialogs contain 235 error segments. Note that according to our definition an error segment can (1) end in either by getting back on track (BOT) with perhaps a complete success, acquiescence or abort (2) be nested within another error segment. Of these 235 segments, 78% got back on track. Figure 3.1 provides the distribution of error segment length (number of turns) in the data. About 80% of these are between 1-9 turns with most of them between 2 to 4 turns. Of these, the average length of the error segments that eventually get back-on-track is 6.7 and those that never recover is 10. From these numbers alone, we do not know whether the length of the unrecovered errors represents something about the system or user, or if it represents some threshold of user tolerance for error resolution beyond which users will simply hang up rather than continue. We present analysis results on the following points: (1) Categories of error perception (2) User behavior under error including user initiative in error vs. non-error situations.

3.3.1 Categories of error perception

Here, we see whether the manner in which the user discovered the error affects the time to get back on track. In the case of a system prompt repetition or a system rejection, the user is explicitly made aware of an “error” (from its perspective). In the case of an implicit confirmation or a system non sequitur, it is up to the user to notice that an error has occurred and draw the system’s attention to this. In Table 3.1, we present error segments grouped by the way in which the user becomes aware of the error, to see if the way in which the error is discovered affects the time to recover or success in recovery. We can roughly divide the error discovery types into high frequency (system rejection, implicit confirmation, & system prompt repeat), and low frequency (explicit confirmation & non-sequitur). Among the high-frequency error discovery types, it is striking that implicit confirmation results in a much longer time to get back on track (10 exchanges vs. 6), and a much lower rate of getting back-on-track at 68%, compared to 80% and 90% for the other high-frequency errors.

Error perception	# of err segments	avg err length for BOT	avg err length not BOT	%BOT
Reject	35	6	7.8	83%
Implicit	25	9.6	14.6	68%
Repeat	21	5.8	13	90%
Explicit	10	5.5	8.75	60%
Non-seq	9	6	7.5	77%

Table 3.1: Lengths of error segments which did get back-on-track (BOT) and those which didn’t, as well as the percentage of errors that eventually got back on track.

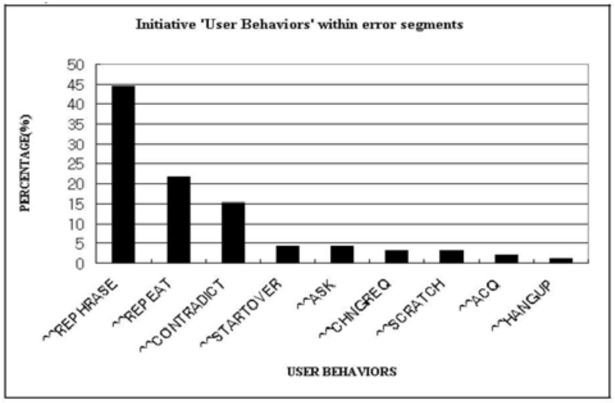


Figure 3.2: “User Behavior” after the first error within an error segment. Rephrasing was the most frequent user behavior and Hang-up was the least frequent user behavior.

3.3.2 User behavior under error

We next examine the distribution of user behaviors in coping with errors. Figure 3.2 shows the distribution on the user behavior immediately following an error (in the previous turn).

The next two tables show the distribution of user strategies for segments that eventually did get back on track and for those that never got back on track:

frequency normalized for length of errors	User strategy in Errors that got back-on-track
0.130	Repeat
0.117	Rephrase
0.077	Contradict system
0.055	Start over
0.045	Ask
0.022	Change request
0.015	Scratch
0.005	Acquiesce to error

Table 3.2: Prevalence of user strategies in error segments which eventually got back on track.

We observe that users in the successful error recoveries (see Table 3.2) use significantly (p ¡ 0.1, ANOVA) more rephrasing than those in the unrecovered errors and less contradictions (e.g. “not 3 am, 3 pm”) (Table 3.3). They also make use of the “start over” and “scratch” features more to terminate error episodes rather than trying to repair chains of errors. Users in successful error recoveries were also much more likely to work around system weaknesses by changing their travel plans. While this apparently got the dialog back on track, it is not a viable strategy for real travel arrangements.

frequency normalized for length of errors	User strategy in Non-back-on-track
0.114	Repeat
0.102	Contradict system
0.071	Rephrase
0.055	Hang up
0.031	Start over
0.024	Ask
0.012	Scratch
0.012	Acquiesce to error
0.004	Change request

Table 3.3: Prevalence of user strategies in error segments which did not get back-on-track.

Degree of Error and User behavior

Errors in spoken dialogs are not merely binary valued and it is critical to incorporate the degree of error into the modeling. To illuminate user behavior under error further, we considered the user response conditioned on the system strategy to estimate the probability $P(UserBehavior||SystemBehavior)$, $P(U||S)$ from now on. It has been well accepted in the field that ASR word error rate (WER) is a good correlate of dialog performance [59]. Hence as a first approximation, we smoothed the probability mass of $P(U||S)$ using an exponentially-weighted WER measure (1-10**(-WER*k/100)) that

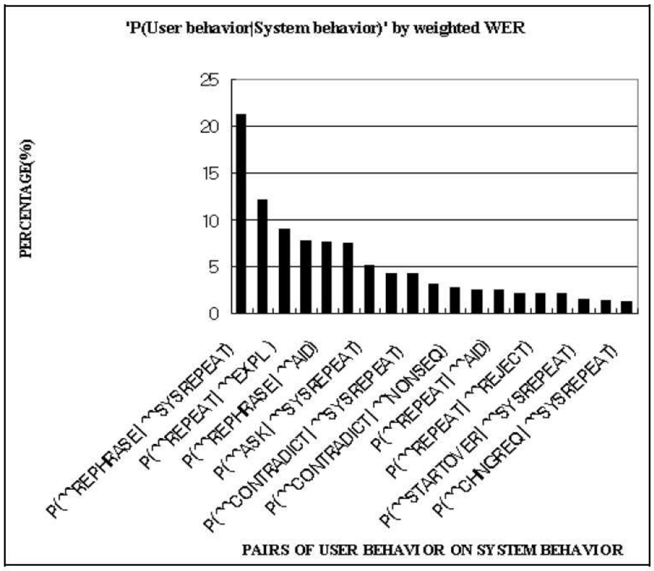


Figure 3.3: P (User behavior — System behavior) smoothed by exponentially weighted WER.

maps WER (which can be between 0 and infinity) to a range between 0 and 1. For the calculations below we chose k=1; it could vary from system to system. The results are shown in Figure 3.3. The most common user behavior here is rephrasing or repeating the previous request, contributing to 82% of all user responses under error. Canceling/changing the previous request or starting over are relatively rare user behaviors under error. This is further exemplified in Figure 3.4 that shows the conditional (smoothed) distribution for $P(U||S = SYSTEMREPEAT)$, corresponding to a highly popular system strategy when the system is "cognizant" of an error.

It is similarly interesting to look at user behavior when the system is not (necessarily) cognizant of an error such as when using an implicit confirmation strategy. Figure 3.5 shows the smoothed distribution for $P(U||S = IMPLICIT)$. Not surprisingly, the user is most likely to contradict the erroneous system behavior.

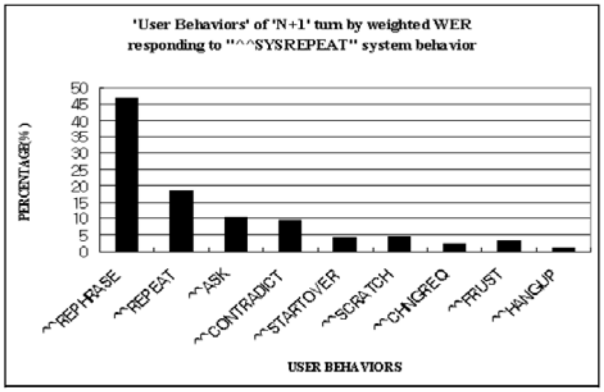


Figure 3.4: Smoothed Conditional Probability for User Behavior in (N+1)th turn based on weighted WER of ’IMPLICIT CONFIRM’ system behavior in the N-th turn.

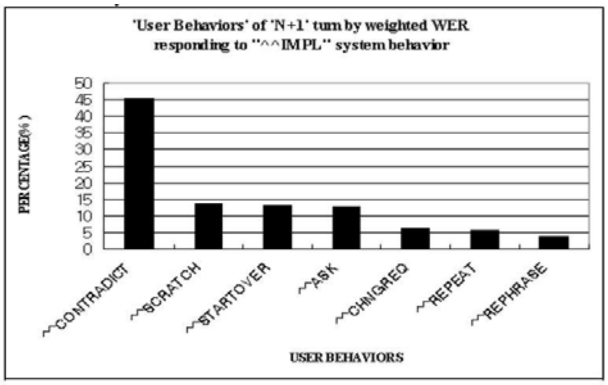


Figure 3.5: Smoothed Conditional Probability for User Behavior in (N+1)th turn based on weighted WER of ’IMPLICIT CONFIRM’ system behavior in the N-th turn.

User initiative in error and non-error environments

Here we look at the user’s tendency to use initiative over the course of the dialog. We have considered user initiative to be the cases where the user did not simply respond to system prompts, but attempted to guide the dialog themselves. The one part of the dialog that often looks the most like user initiative (and which often fails) is the response to the open prompt at the beginning of most of the dialogs. However, since this is a free-form answer to an open question, we have not tagged it as initiative. It is clear from Table 3.4 that user initiative behavior is significantly more in error segments than not (p ¡ 0.05).

User Initiative tag	Frequency in error segments	Frequency in non-error segments
Ask	0.0319	0.0060
Contradict	0.0707	0.0121
General initiative	0.1647	0.0424

Table 3.4: Frequency is normalized over all dialogs.

3.4 Conclusion

Modeling user behavior is one of the most challenging problems in spoken dialog systems research. Empirical analysis and modeling using real user data helps to illuminate user behavior patterns. The analysis reported represents a preliminary attempt at understanding user behavior under error and uncertainty in spoken dialogs. Results show that users discovering errors through implicit confirmations are less likely to get back on track (or succeed) and take a longer time in doing so than other forms of error discovery such as system reject and reprompts. Further successful user error-recovery strategies included more rephrasing, less contradicting, and a tendency to terminate error episodes (cancel and startover) than to attempt at repairing a chain of errors. The most frequent user behavior to get back on track from error segments when the system signals errors is to “rephrase” and “repeat.” When a user discovers an error, say through an implicit confirmation, the user tends to “contradict” or “cancel” the action rather than “rephrase” and “repeat.” There are many open and confounding issues. One key issue relates to incorporating user behavior priors (i.e., probabilities) in the model. For example, we observe that some users seem better able to avoid and/or get out of trouble. The authors of [101] observe that in this specific experimental setup, where the subjects were paid participants with no real stake in successful task completion, some users were simply inattentive or careless. In the process of tagging the transcribed data, we additionally observed that some participants had much more trouble than others getting usable ASR output. Table 3.5 looks at some users who participated in 5 or more scenarios. In Table 3.5, two users, A and B, seem particularly successful. Although they appear to have higher numbers of errors per dialog, this is probably because they did not give up, since they also have the highest rates of recovery with relatively short error episodes. Two other users, C and D, seem the least successful. D has a very low percentage of back-on-track errors, and C seems to experience inordinately long error episodes. When we looked at the strategies these users adopted under error we found that all users tried repeating themselves. However, the less successful users frequently hung up on the dialog or started the dialog sequence over; something that the successful users were less likely to do.

UserID	# of	$Errors/Dialog$	%BOT	Avg length
	dials			of error segment
1	9	1.4	.69	8.5
2	9	1.4	.76	8.9
A	8	2.9	.87	7.8
B	8	2.4	.74	4.9
C	5	1.0	.60	10.2
D	5	1.4	.42	6.0

Table 3.5: Error-proneness in users: % BOT is the percentage of error episodes that got back on track.

These types of prior user information need to be learnt and incorporated into the models. Ongoing work focuses on those questions and how a user model interacts with a system model in an optimization framework.

Chapter 4

Analyzing the Multimodal Behaviors of Users by using Concept Matching Scores

"Concept transfer rate is one of the important metrics to measure system performance and user satisfaction."

We investigate factors related to interfacing a speech-to-speech translation device with multimodal capabilities. We evaluate the efficacy of the interactions using a measure for meaning transfer, we call concept score. We show that employing a multimodal interface improves translation quality, in this study, by 24%. We also show that while some users require perfect representation of what they said in order to allow transfer, others accept concept degradation to some extent, in median up to 20% in our experiments. An appropriate system strategy is required to recognize this behavior and guide users towards optimum performance points. For example, we show that appropriate feedback is required to guide the users in their choices of translation method, as 13% of the choices users made are worse than the alternatives the system provided.

4.1 Introduction

Current speech translation technologies support real time spoken language translation, and are applicable in many areas such as medical services and business meetings. Successful applications include Verbmobil [8], which provided a system for multilingual scheduling, including airline and hotel reservations; Transonics [63], a medical diagnosis tool used by doctors; and MASTOR [24], a multilingual automatic *Speech-to-Speech* (S2S) translation system.

Although there has been intensive research on speech recognition technology [103] and on machine translation [42], studies modeling users in S2S translation systems have rarely been conducted. The need for studying user behaviors has already been demonstrated under spoken dialog systems and it is our hypothesis that similar benefits from user studies can be achieved in S2S systems. For our work we draw knowledge from existing studies such as on system evaluation and multimodal interfaces. Kamm and Walker [39] measured the performance of a spoken dialog system in terms of task success and cost (number of turns). These two factors were utilized for maximizing user satisfaction. Oviatt et al [70] reported that users of a spoken dialog system tend to employ more of the available modalities as cognitive load increases with intensifying task difficulty and communicative complexity. Also, Foster et all [22], in analyzing text prediction models, have argued that by modeling the user they can provide improved text prediction. For S2S translation systems, user studies can cover a vast range of topics, such as language, culture, environment, education, and belief systems [11].

In particular, it is important to study the quality of the transferred concepts through the S2S translation systems, and the level of transfer errors users are willing to accept when using the system. Since a significant part of the machine error stems from speech recognition errors, users are able to gauge in some part the degree of degradation by observing the transcribed text of what they said. Some users are more accommodating to these system errors and still go ahead and accept erroneous speech recognizer output as acceptable for translation, knowing well that they increase the chance of bad concept transfer.

Another important research issue is the design of a flexible human-centric interface for S2S translation systems based on user studies, and evaluating performance gains due to such an interface. Potential system designs include speech-only, speech and text output, speech and text input and output, or can include other modalities such as images, touch screens and pen input etc. It is critical to know the resulting improvements through combination of these modalities under specific conditions. For example in emergency care, one would want very little in the way of device confirmation, visual modalities etc, but instead would prefer a very high accuracy for a very limited number of concepts, while in general practice one may accept a much larger range of modalities to allow for a range of concepts and range of accuracy trade-offs.

Like a human translator, a translation device transfers meaning from and to one language, such as English, to another language, such as Farsi (Persian) [11]. The process is lossy. Vocabulary words and phrases need to be changed to their closest representation in the target language, but will often be remapped to more distant equivalents, and grammar and syntax will also degrade. As a result, the original meaning will be altered at several different levels [52], conveying it sometimes quite closely and sometimes poorly. It is important to measure how well meaning is transferred by translation devices. The existing *text translation* metrics, such as BLEU [76] and NIST [15], scores are based on comparisons of several human translations with system-produced translations using n-gram matching. In the study of this paper, to measure how well meaning is transferred by S2S translation system, we introduce a measure called a concept matching score. This score refers to the number of concepts in a user utterance that is carried over to the machine-produced utterance. We evaluate the performance of the S2S system and its subcomponents in terms of the concept agreement between the input and output according to human annotators.

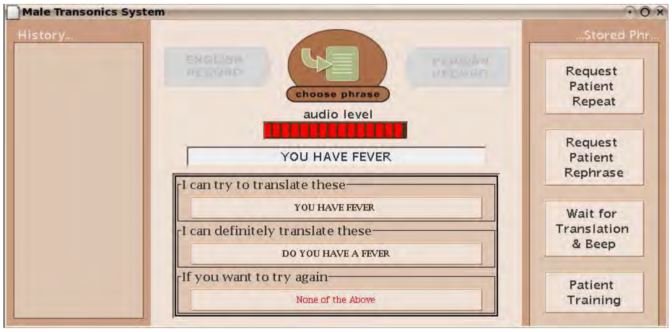
The section is organized as follows. Section 4.2 describes the S2S system used for the experiments and section 4.3, the data collection and annotation. We present results on multimodal versus single modality usage in Section 4.4.1, statistics on user error tolerance in 4.4.2 and analysis on the quality of user choices in 4.4.3. Discussion and conclusions follow in sections 6.6 and 7.

4.2 The Transonics System

Transonics [63] is a speech-to-speech (S2S) translation system, which facilitates two way spoken interactions between English-speaking doctors, and Farsi-speaking patients. This system is aimed at task-oriented interactions in the medical domain.

The *English speaker* (doctor) interacts with the system through two input modalities of audio and a push-to-talk and selection keypad, and receives information through the two modalities of audio and text representations on the screen. In addition there is a more complex direct human-human channel that could potentially encode a lot of information such as gestures and emotions, but that was not very actively used in this collection due to the instructions given to the participants. The *Persian speaker* (patient) interacts with the device only in terms of the audio modality and has no access to the keypad or the screen. The push-to-talk activation for the Persian speaker is handled by the English speaker as well. This asymmetric design allows for minimal knowledge and training of the Persian speaker.

In simple terms, the Transonics design processes the input speech as follows: First, it converts the speech into text (Automatic Speech Recognition – ASR); second, it converts the text into the target language (Machine Translation – MT); third, it plays out the



Male Transonics System

YOU HAVE FEVER

I can try to translate these

YOU HAVE FEVER

I can definitely translate these

DO YOU HAVE A FEVER

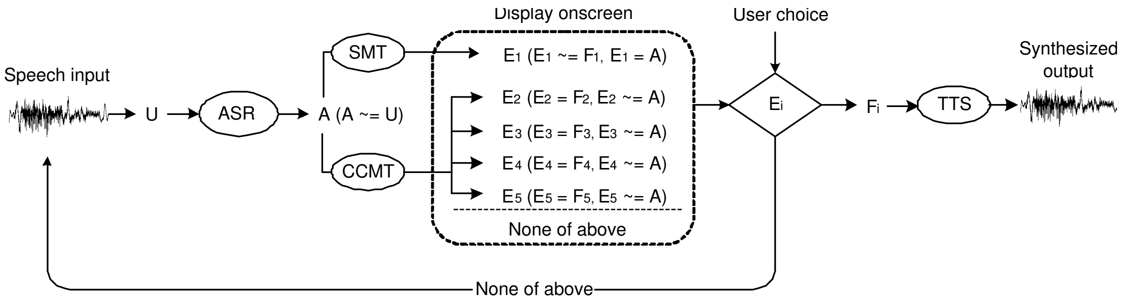
If you want to try again

None of the Above

Figure 4.1: Example image of the system's Graphical User Interface (GUI). After speaking, the English speaker (doctor) can choose one of up to five translation candidates presented onscreen. Section 1 shows the SMT option E_1 labeled with "I can try to translate," while the CCMT options E_i $\forall i \in \{2,3,4,5\}$ are labeled "I can definitely translate these."

translated text (Text-To-Speech synthesis TTS). The MT step operates in one of two modes: The phrase-based translation (often called Statistical Machine Translation – SMT); and the concept based translation (Concept Classification – CCMT). The English speaker sees the various options on the screen after the MT step. We always show one option (E_1) that can be transferred through the SMT path, and up to 4 options ($E_2 - E_5$) that can be transferred through the CCMT path. The CCMT path has the advantage that it provides a very accurate back translation since the concepts known by the CCMT were previously humanly translated. Thus options $E_2 - E_5$ will be transferred very accurately in the target language, while option E_1 will undergo some further channel loss.

Figure 5.3 graphically shows the above description and defines the symbols for subsequent clarity. In short: U is the original user input; A is the ASR belief ($A \simeq U$); $E_1 = A$ is the text that will be translated through the SMT and generate (lossy operation) F_1 ($F_1 \simeq E_1 \simeq U$); and $E_2 - E_5$ is the text already translated and mapped back ("non-lossy", human mapping) into English through CCMT ($U \simeq A \simeq F_i = E_i,\ \forall i = \{2,3,4,5\}$).


Display onscreen
User choice
Speech input
U
ASR
A (A ~= U)
SMT
CCMT
E_1 (E_1 ~= F_1, E_1 = A)
E_2 (E_2 = F_2, E_2 ~= A)
E_3 (E_3 = F_3, E_3 ~= A)
E_4 (E_4 = F_4, E_4 ~= A)
E_5 (E_5 = F_5, E_5 ~= A)
None of above
E_i
F_i
TTS
Synthesized output
None of above

ASR: Automatic Speech Recognition, SMT: Statistical Machine Translation,
CCMT: Concept Classification Machine Translation, TTS: Text-to-Speech,
U: User utterance, A: ASR output, E_1: SMT output in English, E_2 ~ E_5: CCMT outputs in English,
Fi: Farsi translation of Ei(i = 1, 2,3,4, or 5), ~= : Statistical operation, = : Lossless operation

Figure 4.2: The internal procedure of generating speech translation candidates implemented in the Transonics system. A doctor uses two-modality interface (push-to-talk), and sees up to five candidates onscreen; one Machine Translation (MT) candidate (E_1), and up to four Classifier candidates (E_2,E_3,E_4,E_5).

The Persian to English path does not employ this choice interface, but the system has the initiative and selects the best of the 5 options. Due to this asymmetry we will constrain our analysis on the English user behavior.

For example, in Fig. 6.4 we see a screen-shot of the information provided to the English speaker. In this example the speaker said “You have fever?” and sees up to 5 translation candidates (in this case 2) on screen. At this stage the user can detect errors due to the machine speech recognition (ASR) component (option 1) and the ASR and concept classification combined errors (options 2-5). In this case the ASR got the user concept as “You have fever” that in ASR terms is quite accurate but which the translation will likely be a statement. The second option of shows the ASR and concept classification combined, which has resulted in the “Do you have fever” concept. Since concepts are pre-translated by humans, this will result a very accurate, deterministic translation if selected.

4.3 Methods

4.3.1 Data collection

Two participants, a medical professional and an actor-patient, interacted with each other through the S2S device. Both the doctors and patients were monolingual, so communication took place only through the Transonics system. A total of 15 sets of interaction logs were collected from the experiments. The average number of utterances of the English speaker is 33, and that of the Farsi (Persian) speaker is 28. The data was manually transcribed and annotated after the collection. The annotation included concept matching scores between all pairs of same-path utterances such as (A,U), (E_1,U), (E_1,A), (F_1,E_1), (F_1,A), (F_1,U), (E_2,U), (E_2,A), etc. These concept scores were generated by two bilingual annotators that we ensured provided consistent results through training and "calibration" sessions.

The *Concept Matching Score* (CMS) was based on the Linguistic Data Consortium's human assessment metrics [60]. Ma and Cieri [60] say "Adequacy refers to the degree to which the translation communicates information present in the original or in the best of breed translation that serves as a proxy to the original." Based on that the concept matching score compares the number of concepts in an original utterance (source) and the target utterance (destination), either through translation or within the same language, e.g. through a lossy speech recognition channel. The score guidelines for CMS are:

1.0: All concepts are transferred.

0.8: Most concepts are transferred.

0.6: Many concepts are transferred.

0.4: Some concepts are transferred, such that users may sometimes get the whole meaning.

0.2: Few concepts are transferred, such that users rarely get the whole meaning.

0.0: None of the concepts are transferred.

The following example shows a user utterance and its corresponding translation. In this example, the recipient of a translated utterance can easily recognize its meaning even though some of the words in the utterance are not translated correctly. An average concept matching score of CMS[source,target]=CMS[U,F]=0.8 is assigned to the overall system translation path in this example.

Example, Average CMS=0.8:

User (U): DO YOU HAVE DOUBLE VISION

Translation (F_1): VyA SmA v dyd dvgAnh dAryd

(DOES AND YOU HAVE A DOUBLE VISION)

4.4 Results and Analysis

In the study of this paper, we attempt to address three hypotheses. The methods and results are presented in the following three sections 4.4.1, 4.4.2, and 4.4.3.

- **Hypothesis 1:** A multimodal interface, employing both the audio and text modalities, will be better than a single-modality interface utilizing audio only, in terms of translation quality.

- **Hypothesis 2:** Users will accept certain errors from the utterances provided by the system. The degree of degradation in terms of concept representation that different users are willing to accept varies.

- **Hypothesis 3:** Improving the feedback as to when it is appropriate to employ the CCMT path can yield improved translation quality.

4.4.1 Multimodal versus single-modality interface

Users are only able to make choices given a list of available options through the visual modality, and are best able to choose the appropriate option by pen, mouse, or other selection interface. The machine-denoted best choice corresponds in performance to the single modality interface, while the user denoted one to the multimodal interface.

CMS[U,F_u] where $u \in \{1,2,3,4,5\}$ and corresponds to the user choice was compared to CMS[U,F_m] where $m \in \{1,2,3,4,5\}$ and corresponds to the machine-denoted best choice.

To make the comparison more fair we assume that in the case of a single modality a single spoken “yes” or “no” confirmation would be available to the user, emulating in practice the same “None of the above” rejection that the user has in the multimodal interface.

In our analysis, the multimodal interface resulted in CMS[U,F_u]=78% while the single-mode interface produced CMS[U,F_m]=71%. This is a relative improvement of 24%.

By using the multimodal (audio and text) interface, users of Transonics achieved 24% error reduction in translation versus the single-mode interface (audio).

4.4.2 User Error Tolerance

User error tolerance level was measured in terms of the concepts lost between what users said and what they accepted from the utterances that were provided by the system. This is metric CMS[U,E_u] where $u \in \{1,2,3,4,5\}$ and corresponds to the user choice. It is different from the metric in the previous section as it does not consider subsequent device losses in the translation. Note that users can chose to completely reject an utterance, and those rejections are excluded from our analysis.

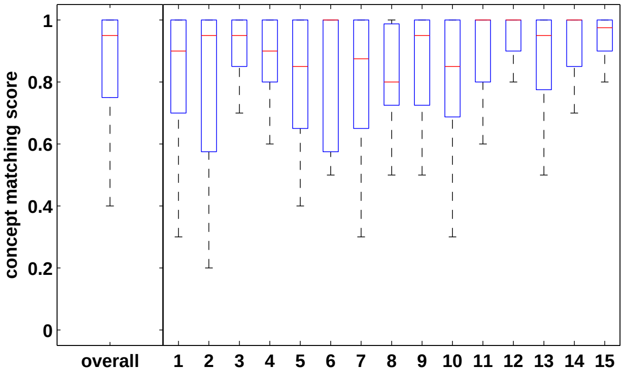


Figure 4.3: A box-plot and concept matching scores of user accepted utterances in the 15 sets of interactions. User retry utterances were not included in the total.

The left part of Figure 4.3 shows a box-plot of concept matching scores of user selected utterances in the 15 sets of interactions. The median score of 0.95 indicates that more than half of the users accepted the onscreen machine-produced utterances when they contained 95% of the concepts in the original utterances. The standard deviation (unbiased) was 0.21, and the mean absolute deviation was 0.17. We also note that users accepted machine-produced utterances with concept matching scores as low as 0.4, which infers quite accommodating users. The mean concept matching score was 0.84, indicating that users on average are accepting of 16% concept loss from the speech recongizer.

Next, we investigated how much users differ in terms of the number of concepts in the original user utterances they would accept when using the Transonics system. The right part of Figure 4.3 shows the box-plots of concept matchings scores of user selected utterances in each set of 15 interactions. Users in the interactions, 6, 11, 12, 14 were picky in accepting machine-produced utterances; half of their accepted utterances have the perfect concept of the original utterances. Users in interactions, 5, 7, 8, 10 were more accommodating than others in acceptance of concept errors in the utterances produced by the system. We observed that some users changed their utterance selections relatively more often than others in terms of the number of concepts accepted, and that some users were relatively consistent in their selections. In the right part of Figure 4.3, the standard deviations ranged from 0.13 to 0.34 in the 15 sets of interactions, indicating that some users were more consistent in their selections (users in interactions 3,12,13,15) than others (users in interactions 1,2,7,10).

4.4.3 The quality of user choices

The quality of user choices using the Transonics system can be measured in relation to the translation quality of the system. As explained in section 4.2 there are blocks of utterance options displayed on the GUI corresponding to two paths of translation: E_1 which will get translated through the SMT, and $E_2 - E_5$, which will be translated through the CCMT.

By investigating user options and corresponding translations, we can measure the quality of user choices; that is: how good were the resulting translations as opposed to user selections.

We measured the CMS[U, E_i] of user selections and corresponding CMS[U,F_i] either provided by the SMT path ($i = 1$) or by the CCMT path ($i \in \{2,3,4,5\}$). Figure 4.4 shows the results. As expected, there is no drop between the two when the CCMT path is the active one, but there is notable drop in the SMT path. This however is counterbalanced by the fact that the E_1 option is on average significantly better than options $E_2 - E_5$, which is also reflected in the 16% higher preference for E_1 by the users.

This analysis motivates us to improve options $E_2 - E_5$ by enriching the concept domain. Despite the degradation in the best-concept classification performance that that would entail, that still is countered by the fact that the user is given a 4-best list and that $E_j = F_j \; \forall j = \{2,3,4,5\}$.

In the above analysis and in figure 4.4 we considered the input of the user as the detected path. When analyzing the SMT path and the CCMT path for the same utterance however, we found that in 13% of the data the users decided to take through the SMT path, performance would have been improved if the users had chosen the CCMT path. In the 13% of the data, we found that the average concept matching score of 0.83 was degraded to 0.62 through the SMT path, but the average concept matching score through the CCMT was 0.75.

We refer to this phenomenon discrepant translation quality. These are cases where the users were focused so much on getting the minimal CMS$[U, E_1 = A]$ error that they ignored their training that instructed them to chose options from the second category ($E_2 - E_5$) if those are acceptable. Thus they would reject accurate paraphrases. Discrepant translation quality occurs because the additional errors made by the statistical method of the SMT procedure are not shown on the GUI. The following notation represents such a case:

$$\text{CMS}[U, E_1] > \text{CMS}[U, E_i] \ \ \forall i = \{2, 3, 4, 5\} \tag{4.1}$$

but,

$$\text{CMS}[U, F_1] < \text{CMS}[U, F_i] \ \ \text{for some } i \in \{2, 3, 4, 5\} \tag{4.2}$$

Our hypothesis, that improving user feedback on when it is appropriate to chose the CCMT path, can yield translation improvements has proven true. To optimize performance, the translation system needs strategies to elicit better user choices onscreen in cases of discrepant translation quality. This requires better self-assessment by the system as to its expected translation accuracy through the SMT and better guidance to the user

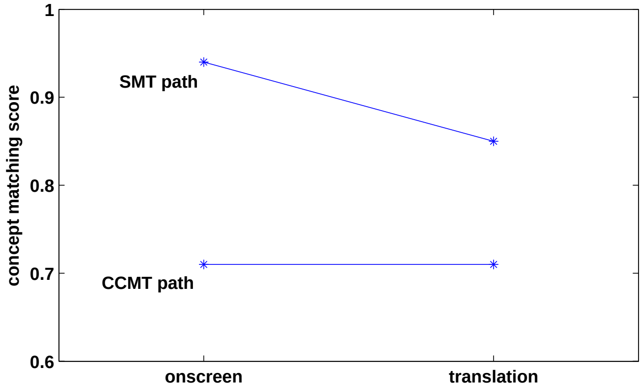


Figure 4.4: Concept matching scores of the onscreen utterances selected by users with the push-to-talk interface of the Transonics, and concept matching scores of the corresponding final translations. User-selected utterances were processed through the SMT path and the CCMT path. SMT is a statistical machine translation and CCMT is concept classification machine translation.

on the expected degradation. To this end, we introduced back-translation functionality in the SMT path in the newer version of the system. Although the effect of this will be studied in future work, preliminary observations seem to indicate that this discourages users significantly more than it was meant to, and encourages complete rejection.

4.5 Discussion

The lessons learnt from the Transonics user studies have been incorporated into the design of our new translation system. We have implemented and are working towards evaluating an agent that will mediate information flow between users and the system in non-intrusive and productive ways. This can be as simple as hinting the user to paraphrase after a rejection to aiding the user in disambiguating ambiguous utterances by providing explanations.

One of the major hindrances in performing these experiments is the difficulty in obtaining consistent annotation of the data. The annotators need to be fluent bilinguals, trained extensively, and calibrated in their responses. In addition this grading of utterances has to be seen in context, and often that is not easy. How do you judge if the speaker would get certain concepts or not? Often that would be impossible to the annotators given that they do not possess the same domain knowledge such as the medical skills of the doctor, and illnesses details as the patient in the medical interaction domain.

An issue in the implementation of real time systems based on this analysis is the correspondence of concept matching scores to real time system confidence levels. We intend to address this in our future work.

4.6 Conclusion

This paper presented an empirical analysis of cross-lingual English-Persian interactions using USC's speech-to-speech translation system. We addressed and validated three hypotheses: First, that additional modalities in the interface aid in communication accuracy, improving relative performance by 24% in the experiments of this study; Second, that users are picky as accepting perfect representations over 50% of the time, and are accommodating as having a median acceptance of 20% concept degradation; and Third that accurate feedback as to the expected degradation of the SMT path can improve the overall translation accuracy of the system, by showing that 13% of user choices led to suboptimal translation.

The design and implementation of useful system strategies to elicit better user choices will be the focus of our future work. In addition we intent to employ other modalities, such as pictures for both input and output, that are universal symbols, to solicit better synchrony among the participants.

Chapter 5

Analysis of Behaviors of Users using a Speech-to-Speech Translation System over Time

“People change over time, which is happening in most cases when using computing systems.”

In the chapter, we report final results from analysis of users who used a multimodal interface of a speech-to-speech translation system during the 4 weeks. Three sets of collected data are investigated for the analysis purpose: user interview data, user survey questionnaire, and log data of the system. In the analysis results of user interview data and survey questionnaire, we report that users incorporated the strategies to cope with system errors in unsuccessful turns, such as repeat, rephrase, change topic, and start over. We also report that users perceived their proficiency in using and learning the system improved during the first three weeks. For the analysis of log data of the translation system, meaning of utterance is considered important in the study. In this regard, we devised a metric called “Concept Matching Score,” which measures the number of concepts transferred from source utterance to target utterance. Using this metric, we first report the distribution of transferred meaning level in two cases; successful and unsuccessful interaction turns of conversations. 91% of utterances in the successful turns contained more than half the meaning of the original user speech, and 90% of utterances in the unsuccessful turns contained less than/equal to half the meaning of the original user speech in our data. Second, we investigate the meaning transfer level by the multimodal interface comparing with that by the speech-only interface. We observed improvement of meaning transfer by 33% and by 11% through the multimodal interface in comparison with two speech-only interface settings respectively; one without and the other with filtering unsuccessful interaction turn. Third, we report that users gradually accepted machine-produced utterances more during the 4 weeks. Further analysis showed that users became more accommodating to the system errors after having experiences of using the system, such as functional word insertion errors, which usually does not impact on the final translation quality. In general, users of speech-enabled interface have a strategy to deal with system errors, which tends to change the length of speech. In our report, the length of user speech increased after successful interaction turn, and decreased after unsuccessful turn. During the 4 weeks, average length of user speech was reduced gradually in the later 3 weeks.

5.1 Introduction

Speech is considered as a promising interface for future computing systems. It enables us to talk freely to/through machine without extra efforts of learning interface functionalities. Among many applications using speech interface, speech translation systems mediating communication between people who speak different languages are demanded as the world become smaller, which is the system of the present study. In the field of text-based translation, there are already many commercialized applications; for example, Google introduced “Google Machine Translation Systems beta” (http://www.google.com/translate_t) in early 2007. In the field of speech translation, commercial applications were yet to be brought in, though research activities are growing; Transonics (now SpeechLinks) system from USC [63], IBM MASTOR system [24], and SRI IraqComm.

To implement a well-performing speech interface for the Speech-to-Speech (S2S) translation system, intensive research efforts are demanded; from speech recognition and translation to cognitive load theory ([103, 42, 67]). In particular, studies about modeling a user are critical for the wide application of the system, and also higher user satisfaction. There are some projects successfully carried out with user information from user modeling community: research and application with user demographic, culture, and preference [86, 49, 37, 35]. Likewise, the speech interface equipped with user information would provide benefits by adapting to diverse people efficiently and expedite conversations between users, and finally users become comfortable using the interface. It can be accommodated to various users in multiple levels of factors, such as gender, error tolerance, and expertise of using the interface.

In practice, most of user modeling studies in the speech technology community were conducted for spoken dialog systems, yet rarely done for spoken mediating systems (e.g., S2S translation system). Few research activities were conducted in parts of the Verbmobil system [98] and in our previous work [92]. Some studies with spoken dialog systems are simply applicable to spoken mediating systems, but it depends on the cases. Important user modeling studies of speech technology community include design and evaluation of multimodal interface [71, 16, 14], analysis of user behaviors [71, 93], probabilistic user model [18, 107], utility-based model [33], knowledge-based model [48], and user simulation [55, 18, 89].

With comprehensive approach using the previous studies, a motivation of the study comes from the perspective that users get used to the system as they use it more. In [57], evolution of expertise over time with desktop application was reported. Conventionally, the expert users are regarded as experienced users with the system [38]. Another motivation is that using multiple input modalities (e.g., speech, mouse, touchpad, and keyboard) together is beneficial for users in many aspects. In the previous work, cognitive load was reduced through a multimodal interface, in comparison with the speech-only interface of spoken dialog systems ([67, 71]). Also, it was reported that the multimodal interface significantly improved user experiences [14]. Indeed, it becomes a general trend to use multiple modalities together to achieve a goal regardless of its usage. As in another study [16], speech-centric multimodal interfaces are growing to be a popular topic in research, though unimodal interface are common at the present time.

In the study, we set up and performed a scenario-based experiment, in which native speakers of English and Farsi interacted with each other using a multimodal interface of a speech-to-speech translation system. In total, three different types of data were collected from the experiment – interview, survey and log data, and we investigate the data in the following aspects: (1) user opinion about the multimodal interface, the system, and the experiment; (2) user satisfaction, and user perception on the level of proficiency in using the multimodal interface, and general speech interface and technologies; (3) user actions upon successful/unsuccessful interaction turn, focusing on user retry/accept behavior and user utterance length; (4) meaning transfer rate through the multimodal interface of the system.

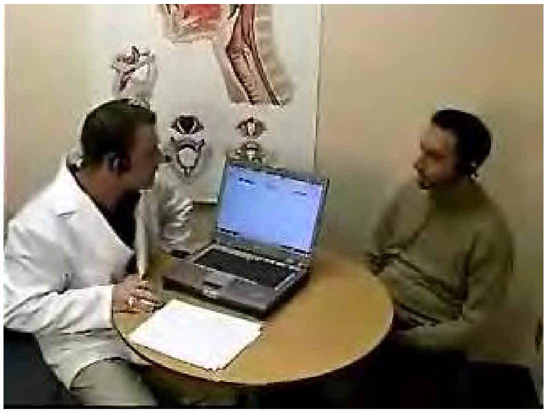
One critical point in the study is that we consider the meaning as part of a metric to assess performance of the translation system. Like a human translator, a translation system transfers meaning from one language, such as English, to another language, such as Farsi (Persian) [63]. The process is inherently lossy. Vocabulary words and phrases need to be changed to their closest representation in the target language. However, they will often be re-mapped to more distant equivalents, and grammar and syntax will also degrade. As a result, the original meaning will be altered at several different levels [52].

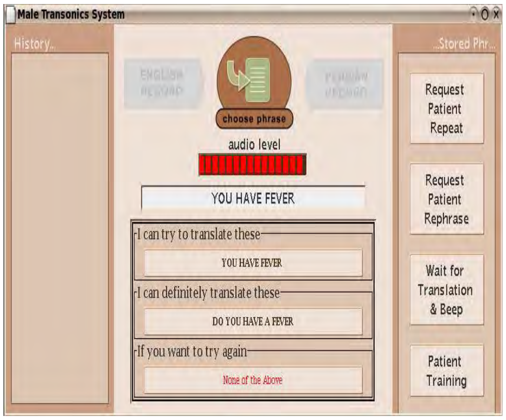
It is conveyed sometimes quite closely and often poorly. It is important to measure how well meaning is transferred by translation systems. Existing text translation metrics, such as BLEU [76] and NIST [15] scores, are based on the comparisons of several human translations with system-produced translations using n-gram matching. In the study, in order to measure how well meaning is transferred by S2S translation system, we introduce a measure called, "Concept Matching Score." This score refers to the number of concepts in a user utterance in the source language that is carried over to the machine-produced utterance in the target language. We evaluate the performance of the S2S system and its subcomponents in terms of the concept agreement between the input and output according to human annotators.

The section is organized as follows. The system used for the experiment is described in Section 5.2. The collected data with its descriptions are laid out in Section 5.3. The results are explained in Section 5.4. The discussion and conclusion are in Section 6.6 and Section 7 respectively.

5.2 System

Transonics system is a two-way translation system with a multimodal interface, called "push-to-talk." Speech and visual are two input modalities for this Graphical User Interface (GUI). The system facilitates two way spoken interactions between an English speaking doctor and a Farsi speaking patient. The goal of the system is to facilitate a task oriented rather than a free-form socio-emotional interaction between two participants. The domain of the system is not fixed, but primarily the medical-related conversations. Figure 5.1 shows a set-up scene of performing a diagnosis for the disease of a patient by a doctor, and one example of recognized user speech with the GUI.





YOU HAVE FEVER

I can try to translate these

YOU HAVE FEVER

I can definitely translate these

DO YOU HAVE A FEVER

If you want to try again

None of the Above

Figure 5.1: A set-up scene of conversations using the Transonics system between English-speaking doctor and Farsi-speaking patient (left). The doctor is conducting a diagnosis of the disease of the patient. The GUI (right) of Transonics shows an example of the recognized results of what user said, “YOU HAVE A FEVER?”

By design, the interface control of the system is asymmetric in the sense that the (English-speaking) doctor has exclusive control over the interface, and access to the GUI, while the (Farsi-speaking) patient does not. This was to allow even untrained and non-educated patients access to the system. Under this asymmetric interface design setting, the monolingual patients are assumed to be untrained in using the system, and to ensure uniform results in the experiments, they are not even 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. If deemed confident, the cluster-normalized form concept will be transferred to the doctor, and if not, 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 repeat, verbally or through gestures, repetitions or repairs. 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 with the user retry option.

With the *push-to-talk* interface, users initiate a speaking turn which has its advantages and limitations; users verify concepts before executing the final decision for ‘speaking-out,’ but work under less spontaneous and less natural environment. In the recognized example in Figure 5.1, when the doctor says, ‘You have fever?’ the user can decide whether to synthesize ‘You have fever’ or ‘Do you have a fever?’, which are translated through machine translation or statistical classifier.

The internal process of the Transonics system involves seven components. Figure 5.2 shows a simplified block diagram of Transonics with its components. 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 form 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 is the center of mediating messages between the modules, and interacts with the MT/classifier, the GUI and the TTS to deliver the data to the user.

To better understand the translation system operation, and the associated issues, we can identify three distinct operations in the process. The first is the conversion from speech (audio) into a textual transcription of the spoken utterance through statistical pattern recognition. This procedure is commonly referred to as Automated Speech Recognition (ASR), and is an inherently lossy operation, i.e. often the transcript may not accurately represent what the user characterized by deletion/insertion/substitution of the spoken words. The second procedure is the machine translation (MT). At this stage the text is mapped from the source language (e.g., English) to the target language (e.g., Farsi). We have two parallel approaches to this step, both again statistical in nature, and

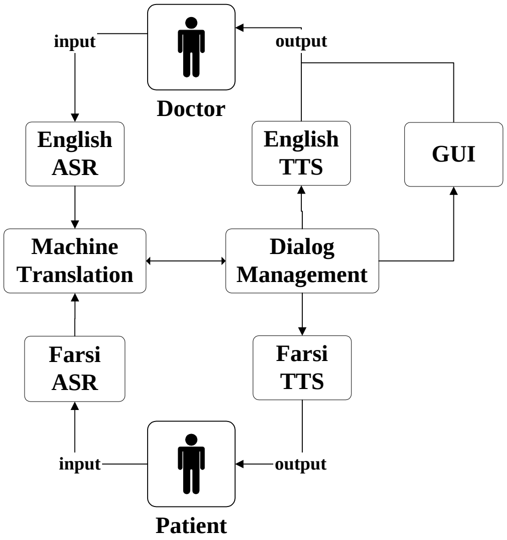


Figure 5.2: Simplified data flow diagram of our two way speech translation system for doctor-patient interactions. English and Farsi Automatic Speech Recognition(ASR) models get the input from users (doctor and patient, respectively) while the Machine Translation(MT) module is responsible for automatic translation and classification of user utterances. The Dialog Manager(DM) manages the interactions between the modules, and delivers the data to users through the GUI. Users finally hear the synthesized output through the the Text-to-Speech (TTS) synthesizer.

represent a lossy mapping process. The approaches we consider are a phrase-based statistical machine translation and an utterance concept classifier – details are in the below paragraphs. The third stage is the conversion of the target language transcript from text to audio by synthesizing the speech, through the TTS.

Figure 5.3 graphically shows the details of above description and defines the symbols for subsequent clarity. As described above, the MT step operates in one of two modes: The phrase-based translation (often called Statistical Machine Translation - SMT); and the concept based translation (Concept Classification - CCMT). The English speaker sees the various options on the screen after the MT step. We always show one option ($E1$) that can be transferred through the SMT path, and up to 4 options ($E2$ - $E5$) that can be transferred through the CCMT path. The CCMT path has the advantage that it

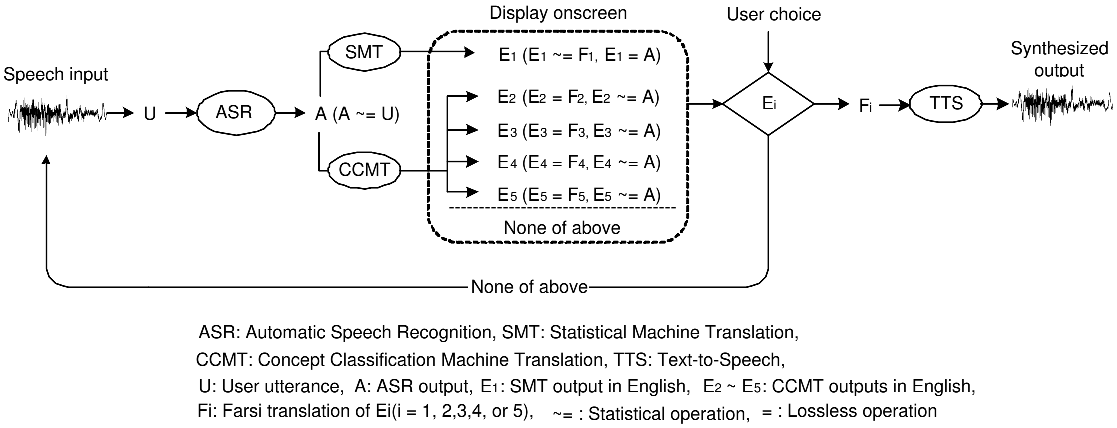

Display onscreen
User choice
Speech input
SMT
Synthesized output
U
ASR
A (A ~= U)
CCMT
E1 (E1 ~= F1, E1 = A)
E2 (E2 = F2, E2 ~= A)
E3 (E3 = F3, E3 ~= A)
E4 (E4 = F4, E4 ~= A)
E5 (E5 = F5, E5 ~= A)
None of above
Ei
Fi
TTS
None of above
ASR: Automatic Speech Recognition, SMT: Statistical Machine Translation,
CCMT: Concept Classification Machine Translation, TTS: Text-to-Speech,
U: User utterance, A: ASR output, E1: SMT output in English, E2 ~ E5: CCMT outputs in English,
Fi: Farsi translation of Ei(i = 1, 2,3,4, or 5), ~= : Statistical operation, = : Lossless operation

Figure 5.3: The internal procedure of generating speech translation candidates implemented in the Transonics system. A doctor uses two-modality interface (push-to-talk), and sees up to five candidates onscreen; one Machine Translation (MT) candidate ($E1$), and up to 4 Classifier candidates ($E2$ - $E5$).

provides a very accurate back translation since the concepts known by the CCMT were previously humanly translated. Thus options E2 - E5 will be transferred very accurately in the target language, while option $E1$ will undergo some further channel loss.

Description of the symbols is follows: U is the original user input; A is the ASR belief ($A \simeq U$); $E_1 = A$ is the text that will be translated through the SMT and generate (lossy operation) F_1 ($F_1 \simeq E_1 \simeq U$); and $E_2 - E_5$ is the text already translated and mapped back ("non-lossy", human mapping) into English through CCMT ($U \simeq A \simeq F_i = E_i$, $\forall i = \{2,3,4,5\}$).

5.2.1 Multimodal Interface of Transonics

In the previous studies, multimodal interfaces were considered flexible, accommodating to large user differences, and supporting rich expressiveness for the user familiarity in modalities [71]. In this regard, the interface of Transonics was designed multimodally to fulfill the requirement of quality translation, accommodating diverse users, and less flimsy with errors.

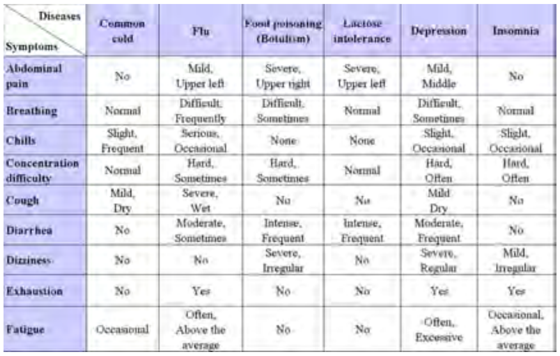
The *push-to-talk* interface of Transonics consists of two input modalities, speech and visual. After voice input, users are able to make choices given a list of available options through the visual modality, and are best able to choose the appropriate option by mouse. Figure 5.3 symbolizes this description. U is the output of user speech, and E_i where $i \in \{1,2,3,4,5\}$ and “None of Above” are the items of the list available to users to choose.

5.3 Data Collection

5.3.1 Experimental Setup

For the experiment, we hired 4 native speakers of English and 4 of Farsi. The age range of the participants was from 20 to 30, and they were graduate and undergraduate students at USC. An hour training session was given to all before the experiment, so they knew what they were supposed to do in the experiment. The training session was about how to do the experiment with given scenarios and the Transonics system. Flash-style interactive instruction was given to the participants for 30 minutes, and verbal explanations were given for 30 minutes. No special training on speech interface was given to the participants in this training session. The compensation for each person was $15 US dollars per hour.

To make more realistic interactions between the participants (participants were not real doctors or real patients), a role-play experiment was designed. English speakers became doctors and Farsi speakers were patients. They were given experimental materials for each role. The materials for doctor-role English speakers were a diagnosis



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

Figure 5.4: Experimental materials for doctor-role English speakers and patient-role Farsi speakers. On the left, a sample of a diagnosis manual of doctor is presented, which is designed for common cold. In the full size table, there are 12 diseases in the column and 30 symptoms in the row. On the right, a patient card for common cold is presented.

manual, a disease treatment manual and a medical term dictionary. The diagnosis manual is a table, consisting of 12 diseases in the column (common cold, flu, food poisoning, lactose intolerance, depression, insomnia, hypertension, high cholesterol, liver cancer, lung cancer, SARS, and diabetes), and 30 symptoms in the row. Each disease's diagnosis manual differs in sorts of symptoms; the 30 symptoms were varied depending on the disease. We developed the diagnosis manual using the medical diagnosis information from **"http://www.medicinenet.com."** The other two experimental materials for doctor are disease treatment manual and medical term dictionary, which are treatment descriptions of each disease and definitions of diseases and symptoms respectively. Farsi speakers were given two experimental materials – symptom card and medical term dictionary. The symptom card is written in Farsi characters, and 4 symptoms of the targeted disease were selected for each interaction session. The medical term dictionary was the same as a doctor's definitions of diseases and symptoms. Example diagnosis manual of doctor and symptom card of patient are presented in Figure 5.4.

5.3.2 Procedure of Experiment

Each conversation team of English and Farsi speakers conducted eight different sessions for 4 weeks, each team performed two interaction sessions per each week. The interaction session was set up with different scenario (different disease with corresponding symptom card), and the doctor worked out to figure out what the disease the patient has with the symptom card. Both parties were not informed about the disease information before the session. All 4 teams followed the same sequence of scenarios: common cold, liver cancer, food poisoning, SARS, hypertension, lung cancer, insomnia and depression. We assumed that the difficulty levels of the eight scenarios were equivalent, which were open-ended. Participants filled out survey questionnaires before and after sessions, and the interaction sessions were video-recorded with a Sony Hi-Fi video recorder for the analysis purpose.

We tried to set up the experiment as objective and in equal conditions as possible. Participants put on headsets and were instructed only to interact through the channel with Transonics, which ensures translations are only being transferred through the device (audio masking effect). The experimenter left the room during the experiment and notified the participants when the time of the interaction session reaches thirty minutes. Therefore, all sessions were finished approximately in thirty minutes.

Figure 5.5 shows a snapshot of one interaction session. The doctor-role English speaker (right) is doing a diagnosis of the disease of the patient-role Farsi speaker (left), under a set-up scenario. The English speaker controls the Transonics system, and there are three experimental materials in front of her. On the other side, two experimental materials of the Farsi speaker are located on the desk.

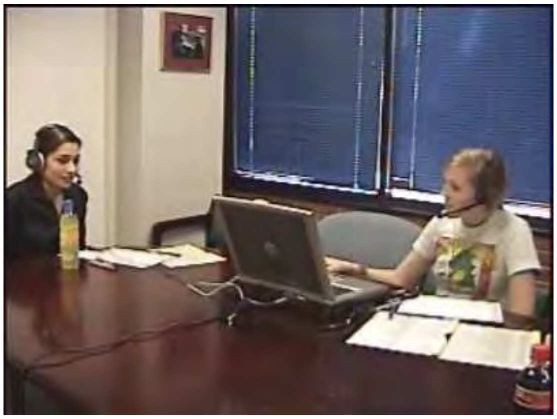


Figure 5.5: A snapshot of an interaction session. A doctor-role native speaker of English (right) controls the device and have a conversation with a patient-role native speaker of Farsi (left) to identify a disease.

5.3.3 Analyzed Data

The data collected during the experiment are in three types: log data, survey questionnaire, and user interview. First, the log data of the Transonics system was collected after each interaction session. It contains all user actions and system information during the session; system confidence levels on the machine-produced utterances, user selections on the machine-produced utterances, recognized hypotheses of the ASR, translated hypotheses of the translation components (SMT and CCMT), recorded user voices, and synthesized system voices in text. Table 5.3.3 shows one cleaned sample of the log data. The system routing tag represents the information flows from the source module to the target module, for example, ‘FADT’ indicates that the data went from the audio server to the dialog management module in text form. In the content column, the processed data are presented, which comes with this routing tag. Second, survey questionnaires were given to the participants before and after the interaction sessions. Also, the initial survey questionnaire was given to the participants to assess the perception on the level of their general technology proficiency, demographic information, and feeling about the

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	ShownAllOptions
FGDT	Choice*1

Table 5.1: 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.

multimodal interface. The questions before the session include feeling of today, the number of experiences using speech interface, and any changes of participants compared to those of the previous sessions. The questions after the session include user satisfaction, perception on the overall system performance, difficulty of topic and using system, and any suggestions. Detail analyses with the survey questionnaire are in the result section. Third, user interview was conducted after each interaction session. In this interview, participants verbally expressed what they felt during the interactions of the session. The experimenter spent 10 minutes for this interview.

5.3.4 Transcription and Annotation of Log Data

The analysis of the log data was two-fold: first, we examined some explicit information in the log data, such as user behaviors (accept and retry), and the machine-produced utterances (speech recognition and translation); second, we annotated the log data with concept scores, which indicate how much correct concepts are transferred by the system, and investigated this annotated log data. The following paragraphs present the details of this second approach.

Concept Matching Score

In [52], meaning is considered as the most important metric for the translation. In this regard, to assess the transferred meaning through the system, we devised a metric, called "Concept Matching Score (CMS)." The idea of the CMS was borrowed from the Linguistic Data Consortium's human assessment metrics [60]. In particular, Ma and Cieri [60] say "Adequacy refers to the degree to which the translation communicates information present in the original or in the best of breed translation that serves as a proxy to the original." Similar to the previous study, the CMS is assigned based on the number of concepts in an original utterance (source) and the target utterance (destination), either through translation or within the same language, e.g. through a lossy speech recognition channel. The CMS scores were manually assigned by human to all the pairs of same-path utterances in the log data, such as (A,U), (E_1,U), (E_1,A), (F_1,E_1), (F_1,A), (F_1,U), (E_2,U), (E_2,A), etc, as shown in Figure 5.3.

To assign CMS scores to utterances of the log data, we hired 4 bilingual speakers, fluent in English and Farsi. Because of a large amount of data (total number of investigated utterances – from English and Farsi speakers – was 2435), the data were divided, two persons took charge in transcriptions of utterances, and the other two in assigning scores to utterances. Two hours of training and calibration sessions were given to these 4 bilingual speakers, in which they were given verbal instructions with some examples for transcribing user speeches and assigning the CMS scores. The CMS scores were assigned based on the following guideline:

1.0: All concepts are transferred.

0.8: Most concepts are transferred.

0.6: Many concepts are transferred.

0.4: Some concepts are transferred, such that users may sometimes get the whole meaning.

0.2: Few concepts are transferred, such that users rarely get the whole meaning.

0.0: None of the concepts are transferred.

Reliability in Concept Matching Score Assignment

We acquired two sets of CMS scores assigned for all the pairs of same-path utterances, which were processed by 2 bilingual speakers. The reason for this was to avoid biased CMS scores. We averaged two CMS scores for each pair of utterances, and used the averaged CMS scores for analysis in the result section. For the reliability in the scores, we computed inter-annotation agreement between two sets of CMS scores. The precision was 0.52, and the number of entries of pairs for the comparisons was 6353.

5.4 Results and Analysis

In the study, we analyzed user behavior and system performance in two types of measures. One is subjective and the other objective. The subjective measure includes user interview and survey questionnaire, while the objective measure includes the analysis of user actions, user utterance length, and machine-produced utterances in the log data. In the analysis of objective measure, humanly annotated Concept Matching Score (CMS) was utilized for the assessment of concept transfer rate from original utterance to target utterance. Statistical analysis toolkit, SPSS 15.0 was used for generating statistical data and graphs.

Before giving details, it may be interesting to see how many diseases the doctor-role participants found out correctly. After each interaction session, we assessed the correctness of the diagnosis result by doctor-role English speakers. They finalized a diagnosis of the disease based on the information collected during the session. Out of 32 interaction sessions, 19 sessions has a correct diagnosis. This means, each participant took part in 8 sessions, and found out the correct disease in 4.75 sessions (std. 0.957) overall. We attempted to investigate user behaviors and its relation to the correctly-diagnosed sessions and the incorrectly diagnosed sessions, but there was no explicit evidence from the results.

5.4.1 Subjective Measure

User Interview

In the interview with participants after sessions, the experimenter asked about what the participants felt about the conversation and the system. Most of the interviewees said they managed to communicate through the system successfully, and during the session, they not only focused on the communication but they thought about the strategies to deal with system errors. The major strategies they used were repeat, rephrase, change topic and start over. In the session of the 4th week, one interviewee mentioned that she became comfortable coping with the errors generated by the system. In particular, some interviewees commented that there were some words unrecognized, and suggested "making the system prone to language would benefit the communication."

Survey Questionnaire

There are three types of survey questionnaire given to the participants. First, every participant filled out a one-time survey questionnaire before the actual experiment. As described in Section 5.3, general questions were given to the participants. In the collected survey questionnaire data, all the participants had no experiences in using speech-enabled systems before the experiment. In this result, the speech-enabled systems represent any applications with speech recognition interface, such as translation or call center spoken dialog systems. Also, the average level of proficiency in general technology (1: comfortable – 7: never comfortable) was 3.0 (std. 1.4) for the 4 native speakers of English, and 2.5 (std. 1.0) for the 4 native speakers of Farsi. The average level of proficiency in dealing with computers (1.0: better than most - 5: worse than most) was 2.25 (std. 0.95), and 1.25 (std. 0.5) in the data of English speakers and Farsi speakers respectively.

In the second and third types of survey questionnaire, which are "before" and "after" survey questionnaires, we investigated only the survey data of doctor-role English speakers. This is because of the asymmetric interface design of the Transonics system, meaning only doctor-role English speakers control the multimodal interface of the system, though still patient-role Farsi speakers use the speech-only interface. The "before" survey questionnaire includes the questions before each interaction session, and "after" includes those after each session. We focus primarily on the "after" survey questionnaire in the following analysis because there was no explicit difference in the analysis results of the "before" survey questionnaire. In the survey data of English speakers, we investigated user perception on overall satisfaction, difficulty in using the system, adaptation to the system, and overall system performance. Table 5.4.1 summarizes overall statistics from the collected user survey data in this regard.

In another question, we measured user perception on their performance on each interaction session – user performance is defined as the level of using and learning the functionalities of the system. In the survey questionnaire, we explained that the user performance can be dependent upon the system performance (speech recognition and

User satisfaction (1: very unsatisfied - 7: very satisfied)	4.6 (0.9)
Difficulty in using the interface (1: difficult to use - 7: easy to use)	5.4 (1.2)
User adaption to using the system (1: difficult to adapt - 7: easy to adapt)	5.3 (1.3)
Overall system performance (1: no concepts delivered - 7: all concepts delivered)	4.4 (0.7)

Table 5.2: Summary of overall statistics from the survey data of doctor-role English speakers, which were collected after each interaction session during the 4 weeks: user satisfaction, user perceived difficulty when using the interface, user adaptation to using the system, and user perceived system performance.

translation), and asked the participants to try to ignore recognition and translation quality of the system when answer this question. This was to reduce the effect of system performance on user performance. Table 5.3 shows what the participants perceived their performance during the weeks in the range of (1: "very bad" - 7: "very good"). The figures show that the participants were getting better in their performance overall, despite a performance drop in the 4th week.

	week 1	week 2	week 3	week 4
perceived user performance	4.25 (0.96)	4.75 (0.96)	5.25 (1.2)	5.0 (0.8)

Table 5.3: User performance during the weeks, in terms of using and learning the functionalities of the system. The answer was in the range of (1: "very bad" - 7: "very good"). Participants were instructed, when they assessed their performance, try to ignore the recognition and translation quality of the system.

Objective Measure

Objective measure in the study is the analysis of the features, such as user behaviors and machine-produced utterances in the log data. Additionally, humanly assigned Concept Matching Scores between all the pairs of utterances were utilized for this analysis purpose. Before going details, it is worthwhile to take a look at general statistics about the data. We present the statistics of the data from doctor-role English speakers, which contains more valuable data for the analysis because of the asymmetric interface design of the system: total number of utterances of 4 speakers during the 4 weeks was 1489; the average number of utterances per interaction session was 46.5 (std. 16.7); the average user acceptance rate in machine-produced utterances per each interaction session was 0.64 (0.09); the average number of words per utterance was 5.16 (2.0); and the average concept matching score between the user speeches and the user-accepted utterances (from the machine-produced utterances) was 0.84 (0.19). The general statistics about the data of Farsi speakers were: total number of utterances of 4 speakers during the 4 weeks was 946; the average number of utterances per session was 29.6 (11.8); the average number of words per each utterance (translated word by word in English) was 3.34 (2.3); and the average concept matching score between the user speeches and the translations (machine-produced) was 0.56(0.39). Note that Farsi speakers did not have an access to the push-to-talk interface, therefore most of utterances were transferred through the system without being filtered by Farsi speakers.

* Distribution of Transferred Meaning from User Speech to Successful and Unsuccessful Machine Recognition.

In the study, users of the multimodal interface accepted or retried the machine-produced utterances depending on certain conditions. The conditions involve the number of concepts of user speech transferred through the system: users speak and see the recognized utterances by the system using the multimodal interface. Intuitively, users accept the machine-produced utterances when “most” of the concepts in user speech are in the machine-produced utterances, and retry with “low” concept transfer in the machine-produced utterances. The question is how many concepts “most” represents when users accept the machine-produced utterances. For example, “most” can represent the perfect concept transfer or half of concept transfer from the original user speech.

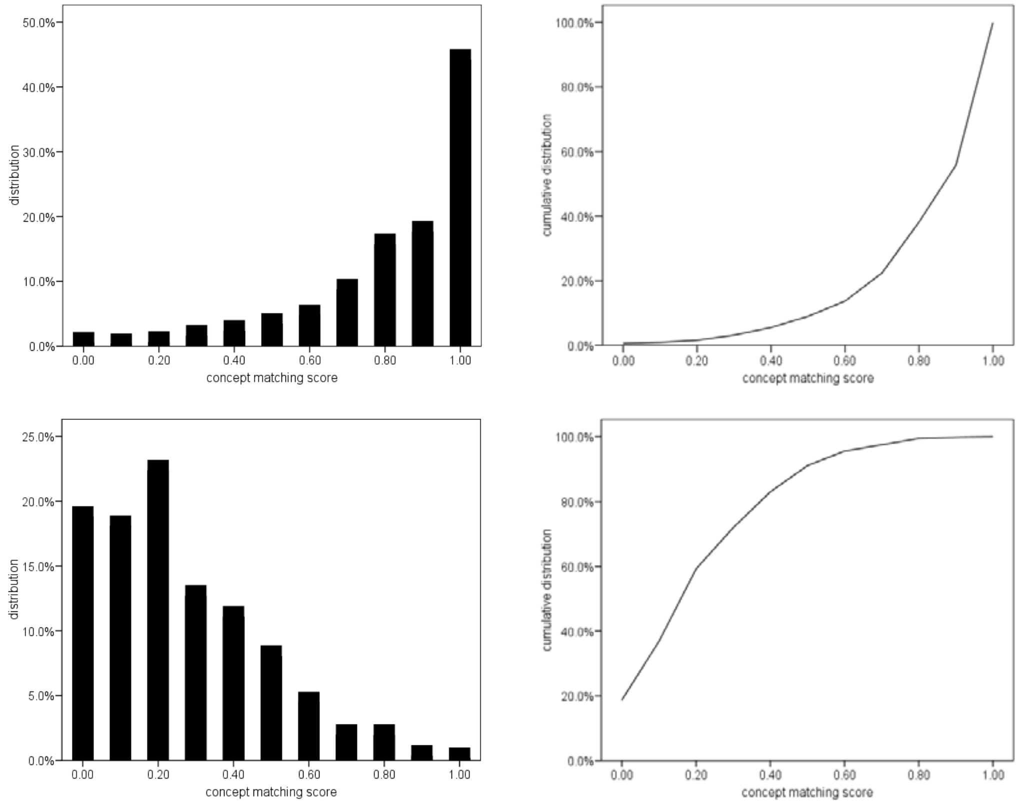


Figure 5.6: Top left: CMS between user speech and user accepted utterance (between U and E_i where $i \in \{1,2,3,4,5\}$ in Figure 5.3); Top right: cumulative CMS of the left distribution; Bottom left: CMS between user speech and user rejected utterance (between U and E_i where i = *None of above*); Bottom right: cumulative CMS of the left distribution.

In this regard, we investigated the number of concepts transferred from user speech to user-accepted utterances (successful turn) and user-retried utterances (unsuccessful turn). Figure 5.6 shows the distributions of Concept Matching Scores in two conditions (successful and unsuccessful), represented by the symbols in Figure 5.3, (U,E_i) where $i \in \{1,2,3,4,5\}$ and (U,E_i) where $i =$ *None of Above*, with the cumulative sum graphs of the Concept Matching Scores.

In overall, CMSs between user speech (U) and user-accepted utterance (E_i, where $i \in \{1,2,3,4,5\}$) and between user speech (U) and user-retried utterance (E_i where $i =$ *None of Above*) were 0.84 (std. 0.19) and 0.25 (std. 0.21) respectively. The individual average CMSs of 4 users between user speech and user-accepted utterance were 0.83 (0.24), 0.86 (0.15), 0.80 (0.2), and 0.87 (0.14) and between user speech, and CMSs between user speech and user-retried utterance were 0.33 (0.25), 0.20 (0.18), 0.29 (0.22), and 0.21 (0.16) respectively.

* Boosted Concept Transfer Rate by Multimodal Interface.

The Transonics system incorporated a multimodal interface with speech and visual input modalities (called "push-to-talk") to expedite higher concept transfer rate in conversations between speakers. The motivation of designing the multimodal interface of a S2S translation system was drawn in parallel with advances in the previous studies ([71, 14]), in which overall system and user performance enhancement were achieved by multimodal interface of spoken dialog systems. In the study, we conducted an experiment comparing concept transfer rates in two settings: (1) multimodal interface; (2) unimodal interface. The log data of the Transonics system were utilized in this study.

Detail experimental settings are the following. When using the multimodal interface, users selected one of the machine-produced utterances onscreen or retry it using "*none of the above*" retry option. After this, the corresponding translations of the user-selected utterances were through the internal procedure (refer to Figure 5.3), and synthesized to the other user. In the unimodal interface setting, the system internally selects the best utterance matched with user speech and synthesizes it (no retry option). We compared the Concept Matching Scores between user speech and translation in these two settings. As shown in Table 5.3, 33% relative CMS improvement was achieved in the multimodal interface setting when there was no filtering option (retry) for bad machine recognition in the unimodal interface setting. To make the comparison more fair, we assume that in the case of the unimodal interface, a single spoken "yes" or "no" confirmation would be available to the user, emulating in practice the same "*none of the above*" rejection that the user has in the multimodal interface. This experiment can be formalized using the

	CMS between user speech and translation	
	without retry option	with retry option
unimodal	0.51 (0.33)	0.63 (0.25)
multimodal		0.67 (0.23)

Table 5.4: Concept Matching Score (CMS) (standard deviation) between user speech and translation in two settings: unimodal interface and multimodal interface. The unimodal interface with retry option was set up with a single spoken "yes" or "no" confirmation which would be available to the user emulating the "*None of the above*" retry option of the multimodal interface.

symbols in Figure 5.3; CMS[U,F_u] where $u \in \{1,2,3,4,5\}$ and corresponds to the user choice was compared to CMS[U,F_m] where $m \in \{1,2,3,4,5\}$ and corresponds to the machine-denoted best choice respectively. In this experiment, we acquired 11% relative CMS improvement in the multimodal interface setting (Table 5.3). Note that, in this setting, only 10% of the utterances from both interface settings were different, which primarily boosted the concept transfer rate in the multimodal interface setting.

* Increase in User Acceptance Rate over Time.

During the experiment, the users using the multimodal interface of the Transonics system accepted or retried the machine-produced utterances onscreen. According to the investigation on the number of user acceptance in the machine-produced utterances, we observed that users' acceptance rate increased during the 4 weeks in overall, as shown in Figure 5.7. User acceptance rate is defined as:

$$\frac{\text{the number of user-accepted utterances in a session}}{\text{the number of whole utterances in a session}} \tag{5.1}$$

This increasing user acceptance trend during the weeks may be related to some factors. Table 5.5 presents the Concept Matching Scores of two cases during the 4 weeks: (1) CMS between user speech (U) and ASR output (A); (2) CMS between user

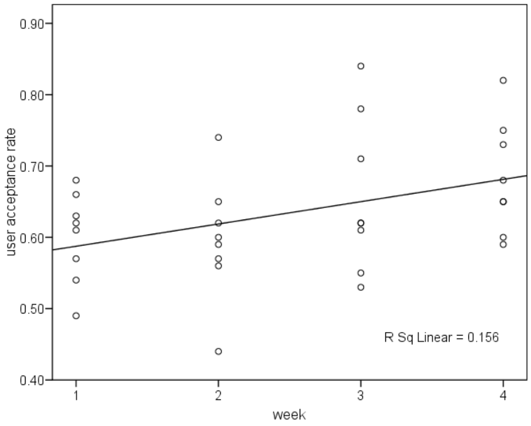


Figure 5.7: Linear trend representing increasing user accepted rates during the 4 weeks. User accepted rates were acquired over the 32 interaction sessions, in which the English speakers accepted the machine-produced utterances using the multimodal interface of Transonics. Each circle represents user accepted rate per each interaction session.

speech (U) and user-accepted utterances in the machine-produced utterances (E_i where $i \in \{1,2,3,4,5\}$). Statistical analysis with ANOVA measure on the first case (CMS between user speech (U) and ASR output (A)) confirmed no significant difference during the 4 weeks: F=2.0, p=0.12. However, ANOVA measure on the second case (CMS between U and E_i where $i \in \{1,2,3,4,5\}$) confirmed a significant difference over the weeks: F=4.17, p<0.01. Post-hoc test with Tukey HSD on the second case confirmed that the CMS of Week 1 and that of week 4 are significantly different. Conjecture is that the reason for the increased acceptance rate during the weeks is that users accepted more functional word errors as they became accustomed to the usage of the system. The example "AND ANY OTHER SYMPTOMS" was observed in our data.

One question is that decrease in concept transfer from user speech to user-accepted utterances (CMS between U and E_i where $i \in \{1,2,3,4,5\}$) may or may not affect on the translation quality (CMS between U and F_i where $i \in \{1,2,3,4,5\}$). In this regard, we investigated the translation qualities during the 4 weeks: 0.69 (0.22), 0.64 (0.22), 0.66

	week 1	week 2	week 3	week 4
CMS (U, A)	0.66 (0.31)	0.64 (0.32)	0.68 (0.3)	0.64 (0.37)
CMS (U, E_i where $i \in \{1,2,3,4,5\}$)	0.88 (0.19)	0.85 (0.2)	0.84 (0.18)	0.82 (0.2)

Table 5.5: Overall CMS (standard deviation) of two cases in the interaction sessions during the 4 weeks. The first case is between user speech (U) and ASR output (A), and the second is between user speech (U) and user-accepted utterances in the machine-produced utterances (E_i where $i \in \{1,2,3,4,5\}$).

(0,21), and 0.65 (0.24). The ANOVA on the translation qualities during the 4 weeks confirms that there is no significant difference (F=1.3, p=0.17).

* Effect of successful and unsuccessful interaction turn on the length of user utterance.

Users using speech interface have an effective way of dealing with system errors. Users reduce the length of their speech and rephrase the previous utterance after system error or in the consecutive chains of system errors. On the contrary, they attempt to speak long sentences when the system works fine. Table 5.6 shows the average utterance lengths of the participants in three conditions: (1) in overall; (2) after accept behavior; (3) after retry behavior. It indicates that the participants spoke relatively longer utterance after accepting previous machine-produced utterance, and shorter utterance after retrying the previous machine-produced utterance.

From an another perspective, we hypothesized that the users of the multimodal interface gradually felt and learned how to deal with system errors effectively by modifying the utterance length while retrying or accepting previous machine-produced utterances. In particular, from our data, users reduced the length of their utterances when error happened, and this trend increased over the weeks. Table 5.7 shows the percentages of reduced utterance length after user retry behavior, through the multimodal interface of Transonics during the 4 weeks. The statistics were from the 32 interaction sessions of

	Overall	After accept	After retry
Average utterance length	5.16 (2.0)	5.26 (2.1)	5.0 (1.9)

Table 5.6: Average utterance length (standard deviation) in three conditions: in overall, after accepting previous machine-produced utterance, and after retrying previous machine-produce utterance. The utterance length is defined as the number of words in an utterance. The statistics were collected from the 32 interaction sessions of the English speakers who controlled the multimodal interface of the Transonics system.

the English speakers. Users in the third and the fourth weeks have higher percentages in reducing their utterance length after retrying the previous machine-produced utterance, compared to the first and second week (though the percentage in the second week did not increase from the first week). On the other hand, the percentages of increased utterance length after accepting the previous machine-produced utterance during the 4 weeks were 62%, 62%, 67%, and 68% respectively (the same utterance length inclusive).

week 1	week 2	week 3	week 4
23%	20%	28%	30%

Table 5.7: Percentage of reduced utterance length after user retry behavior using the multimodal interface of the Transonics system. The statistics were from the 32 interaction sessions of the English speakers during the 4 weeks.

Investigating details about individual difference in the utterance length, ANOVA measure confirm that there is a significant difference in the lengths of individual utterances: F=48.7, $p<0.01$ (English speakers), and F=90.7, $p<0.01$ (Farsi speaker). Table 5.8 shows the difference of utterance length between users. User 1 and 4 (from both English and Farsi) used shorter utterances compared to user 2 and 3. In an attempt to investigate the different utterance lengths of users during the 4 weeks, for example

	User 1	User 2	User 3	User 4
English speaker	4.1 (1.6)	5.7 (2.3)	5.6 (2.0)	4.6 (1.2)
Farsi speaker	3.0 (1.8)	4.4 (2.3)	3.7 (2.3)	1.7 (1.3)

Table 5.8: Average utterance length (in words) and its standard deviation of English speakers and Farsi speakers, which were collected from the 32 interaction sessions during the 4 weeks.

users may have a tendency of decreasing utterance length during the 4 weeks, but we could not find a significant difference in the result.

5.5 Discussion

One of primary hypotheses for the study was that translation quality would improve as users became used to using the system. Intuitively, users become proficient over time in dealing with the system gradually, and they got better system performance eventually. However, the study showed that there was no significant improvement or retrogression in the translation quality during the 4 weeks in terms of correctly transferred concepts. Conjecture is that the reason why translation quality did not improve during the weeks: (1) Too many unknown and combined factors caused system errors (speech recognition and translation errors), which can be mistaken user behaviors, mismatches of acoustic and lexical models between user speech and the system, and intrinsically statistical property of the system. (2) The system could not process all the utterances of users, some words in user speech were not in the vocabulary or in the n-gram matches (language model) of the system. The problem is that this “malfunctioning” (above two conditions) can happen anytime. In the interaction sessions during the 4 weeks, participants worked with given scenarios, which were pretty much open-ended. Therefore, the participants could speak any type of utterances in the domain of the scenarios. In this condition, the “malfunctioning” problem happens in any week during the experiment, not depending on user skills or user experience levels.

Another important issue in the study was bias in the assignment of Concept Matching Scores between utterances. Concept Matching Score (CMS) is a subjective metric assigned by human. Although we hired 4 bilingual people to make two sets of scores, gave careful instructions to them, averaged the scores between the two sets, and investigated statistically huge enough data, still it does not guarantee perfectly unbiasedness in the data. In the future work, we plan to devise more unbiased metrics than CMS for the assessment of the number of transferred concepts from source utterance to target utterance.

5.6 Conclusion

In the study, we investigated three sets of user data to identify various aspects of users using a multimodal interface of a speech-to-speech translation system during the 4 weeks. The various aspects of users include how users react on system errors and how well users use and learn the multimodal interface during the 4 weeks. The analyzed data include interaction data between English speakers and Farsi speakers, user survey questionnaire, and user interview data. In the analysis results of subjective measures, which are user interview and survey questionnaire, it was reported that users utilized some strategies to cope with system errors, such as repeat, rephrase, change topic, and start over. Also, it was reported that perceived user performance increased during the first three weeks, despite a drop in the week 4. For the analysis of the log data, we transcribed user speeches and assigned Concept Matching Score (CMS) between all the same-path utterances, with the help of 4 bilingual speakers. The CMS was used for measuring the number of concepts transferred from source utterance to the target utterance.

As for the findings, first we presented the distribution of CMS over the utterances. 91% of successfully recognized utterances by the system has more than half the concept of user speech (CMS: 6.0–10.0), and 90% of unsuccessfully recognized utterances contain less than/equal to half the concept of user speech (CMS:0.0–0.5). Second, we presented how much improved concept transfer rate we acquired through the multimodal interface of a speech-to-speech translation system. We observed improvement of concept transfer rate by 33% and by 11% through the multimodal interface in comparison with two speech-only interface settings respectively; one without and the other with filtering unsuccessful interaction turn. Third, gradual increase in user acceptance rate over the utterances was observed in our data. Further analysis showed that users became accommodating to the system by accepting more errors, such as functional word insertion error, during the 4 weeks. Fourth, user utterance length increased after accepting the previous utterance (successful turn) and decreased after retrying the previous utterance (unsuccessful turn). Further analysis showed that a trend of increasing percentage of reduced utterance length (after retrying the previous utterance) was observed in the later three weeks.

Chapter 6

User Modeling in a Speech Translation driven Mediated Interaction Setting

“Dynamic Bayesian network is one of the best tools representing user behaviors under uncertain conditions.”

The study addresses user behavior modeling in interactions between two people that 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. Additionally, in a face-to-face scenario, there may be information directly exchanged between the interlocutors, typically through non-verbal gestures. 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 – namely *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 one of the 3 user types is clearly identified as a user keeps his/her consistent behavior in a given interaction condition. In the online experiment, agent feedback is presented to users according to the user types. The analysis showed high user satisfaction and interaction efficiency in the data of user interview, recorded video, survey questionnaire and log of the system. Additionally, we investigate communication patterns in the direct “interpersonal” channel, focusing in patterns on the users’ utterance length. The analysis showed that the average utterance length of a user reflects specific user types and can be in turn used for facilitating interpersonal adaptation. Speech Accommodation Theory supports the argument that greater degree of utterance length accommodation is related to higher user satisfaction in Human-Human interactions.

6.1 Introduction

Spoken conversations have been recognized as the primary information delivery 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 institutions [64, 105, 80, 7]. 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. User modeling has been attempted at different levels and using a variety of approaches. Rich [85] has proposed a 3-dimensional space to describe

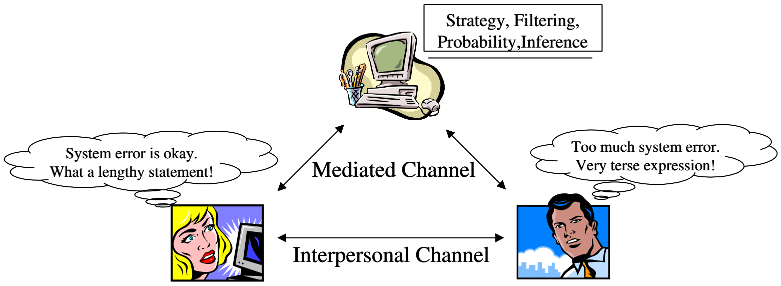


Figure 6.1: There are two channels of parallel interaction between two users: an *interpersonal* path, that contains direct cues such as prosody, gestures, facial expressions, as well as indirect such as adaptation to one others speaking style and a computer *mediated* path containing mainly translated lexical information.

the relationship between user models, defined as the knowledge about people, and their uses. In Table 6.1 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.

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

Table 6.1: User model dimensions(Dimension 1,2,3) based on the knowledge about people [85].

While there has been a fair amount of excellent user modeling work in the context of human-machine spoken dialogs including user simulation [17, 25], reasoning about a user's goal or intention [30], user expertise modeling [47], and evaluation techniques [58], 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 building a user model. The profile acquired from a user can be used for generating an appropriate system response, such as personalized search [78], or in appropriate help to the user when needed [30, 6, 102]. For the objectivity and extendability of the system, we prefer to use predictive statistical user models. It is considered a powerful approach to model user behavior [108] and its effectiveness has been demonstrated by previous research [47, 50]. We specifically propose incorporating 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 (Figure 6.1) – 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 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 one others speaking styles). Our analysis in this section is restricted to aspects of the verbal behavior in these channels.

The rest of the section 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 6.2, in Section 6.3 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 6.4, online experiment with agent feedback is presented and results are reported. Motivated by Speech Accommodation Theory

(SAT), in Section 6.5 we explore verbal (lexical) patterns in the direct human-human, interpersonal channel. A discussion of the results and future directions is provided in Section 6.6. It includes a discussion about the relation between the user behavior in the mediated and the interpersonal channels, as well as a preview into the design of a cross-lingual conversation assistant. Finally, conclusions as well as a description of future work plans are given in Section 6.7.

6.2 System and Dataset

6.2.1 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 [64]. 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 socio-emotional 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. It is 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

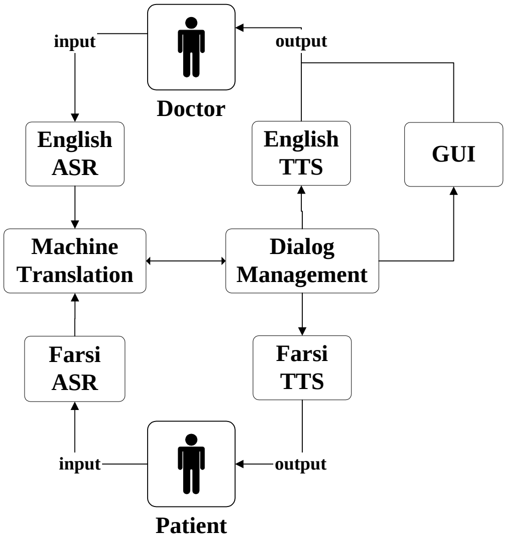


Figure 6.2: Simplified data flow diagram of our two way speech translation system for doctor-patient interactions. English and Farsi Automatic Speech Recognition(ASR) models get the input from users (doctor and patient, respectively) while the Machine Translation(MT) module is responsible for automatic 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).

satisfaction is one of the most important efficacy metrics of medical domain interactions [29, 87].

A functional block diagram of the system and its data flow are shown in Figure 6.2. 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 form 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

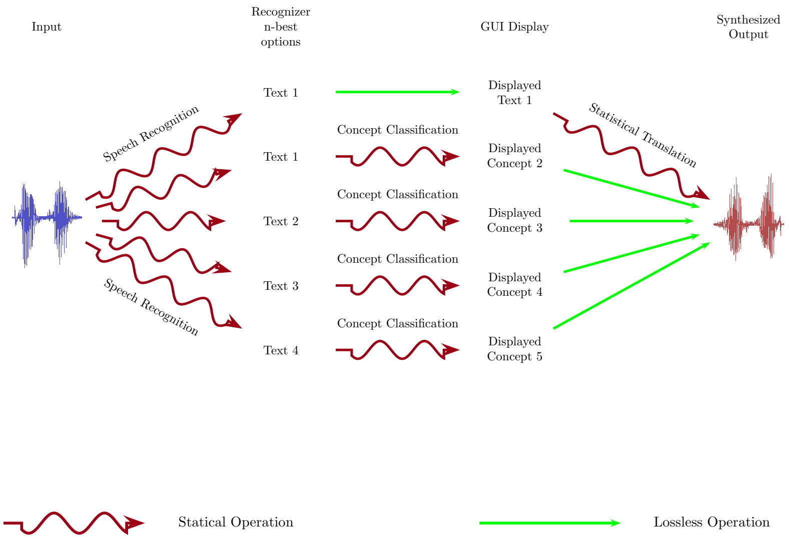


Figure 6.3: The internal procedure of generating speech translation hypotheses in our system. Two parallel mechanisms are implemented. In the first one, the topmost recognition candidate i.e., the first-best choice of the ASR – that has already gone through a lossy speech to text mapping process – will 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.

by the GUI is made available only to the (English-speaking) doctor, who is assumed to have the primary control of the interaction.

To better understand the translation device operation, and the associated issues, we can identify three distinct operations in the process. The first is the conversion from speech (audio) into a textual transcription of the spoken utterance through a statistical pattern recognition. This procedure is commonly referred to as *Automated Speech Recognition* (ASR), and is an inherently lossy operation, i.e. often the transcript may not accurately represent what the user characterized by deletion/insertion/substitution of the spoken words. The second procedure is the translation. At this stage the text is mapped from the source language (e.g., English) to the target language (e.g., Farsi).

We have two parallel approaches to this step, both again statistical in nature, and represent a lossy mapping process. The approaches we consider are a phrase-based statistical machine translation and an utterance concept classifier. 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.

By design, the interface control of the system is asymmetric in the sense that the (English-speaking) doctor has exclusive control over the interface, and access to the GUI, while the (Farsi-speaking) patient does 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 Figure 6.4 and the hypotheses belong to one of two classes:

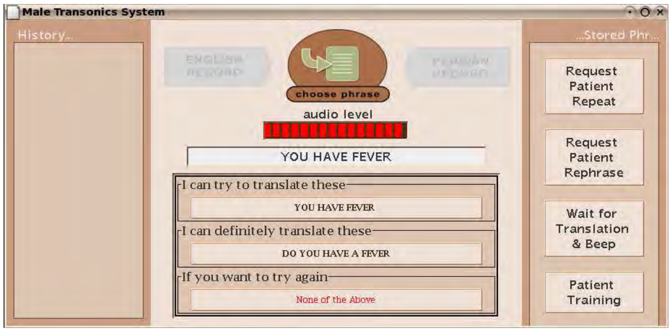
1. The first is the English transcription of what the machine thinks the user said. The machine does not provide a translation on the screen (presumably it would not be useful for the doctor who doesn't know Persian) but a statistical phrase based translation would be provided to the patient if the doctor chooses this option. However, such statistical machine translation *can not* 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 reducing 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 over 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 translation steps, and that the translation the patient will hear will be lexically identical to the hypothesis displayed on the screen. Figure 6.3 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 Figure 6.4 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 even 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. If deemed confident, the cluster-normalized form concept will be transferred to the doctor, and if not 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



YOU HAVE FEVER

I can try to translate these

YOU HAVE FEVER

I can definitely translate these

DO YOU HAVE A FEVER

If you want to try again

None of the Above

Figure 6.4: Transonics system screen GUI. After speaking, the user(doctor) can choose one of several hypotheses presented on the GUI.

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 6.2. These results stem from 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 the Table 6.2 the word error rate(WER[1]) and the IBM BLEU[2] scores are provided.

[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 $0 \leq$ IBM BLEU ≤ 1, even the best human expert translators can only achieve average ranges of near a half of that.

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%

Table 6.2: DARPA evaluation on medical domain for the speech translation system of this paper. Component and Concept measures as: 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.

6.2.2 Data-set

The data analyzed are from 15 interactions between doctors and standardized patient actors. 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 procedure information.

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 transcriptions. It may be interesting to note some relevant information regarding the data characteristics. The average number of turns (each turn is 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

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	ShownAllOptions
FGDT	Choice*1

Table 6.3: 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.

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 hours.

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.

6.3 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 [79, 61] and stereotypes [84]. Following these generally classified assumptions, considerable research efforts have been undertaken covering various topics and systems: Komatani [47] introduced a general user model with skill level, knowledge level, degree of urgency in a spoken dialog system, Carberry [9] modeled user preferences in a natural language consultation system, Conati [12] proposed how to manage uncertainty in a student model by performing assessment and recognizing plans for a tutoring system, and Prendinger [81] 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 [46, 74, 97].

Error handling mechanism is important in the design and optimization of a spoken dialog system. The spoken communication channel between a human and a machine is inherently noisy, and can further be exacerbated by user-dependent uncertainty such as due to limited world or task knowledge. The significance of considering user behavior under problematic conditions in human-machine interaction is demonstrated for example by our prior work [**?**], where we highlighted the importance of repeating and rephrasing cues. Similarly, the work of Batliner [5] 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 with the motivation that distinct interface strategies can be developed for each case. 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. 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 similarly to what previous studies have demonstrated [36, 40, 47].

6.3.1 Analysis of Repeat/Rephrase(“Retry”) Behavior

Repeat or rephrase (Retry) is 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 observed retry behavior. In addition to the small variance in the speech recognition error, we observed that most of errors stem from insertions

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

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.

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’s accommodation to speech recognition 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 [47], user skill level is defined by the maximum number of slots filled by utterances and in [43, 12], 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 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 each user’s retries. We assume that accepting different ranges in WER depends significantly on the user type, as conceptualized in Figure 6.5, and hence we define

- *Accommodating*: users tend to accept highly erroneous transcriptions compared to other users.

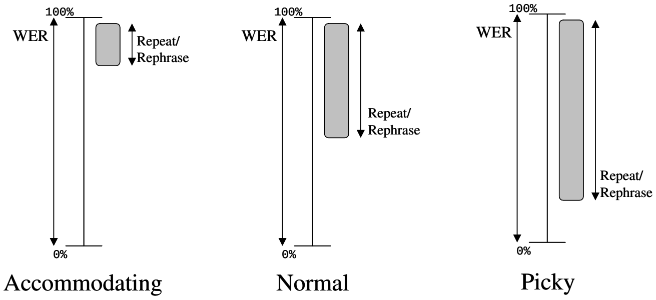


Figure 6.5: The *Accommodating* user tends to "Retry" significantly less than the other users while the *Picky* user tries significantly more. A user in between these extremes is defined to be a *Normal* user. WER is the speech recognition Word Error Rate and the above graph semantically demonstrates the ranges of WER for which each user type tends to "Retry."

- *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 3 classes as shown in Figure 6.6. 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

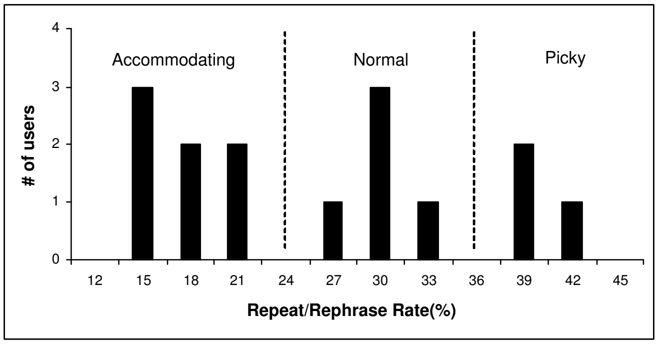


Figure 6.6: The quantized retry rate over 15 interaction sessions on the doctor side. The criteria (average retry rate) based on the data analysis led us to categorize the users into 3 types: Accommodating, normal, and picky.

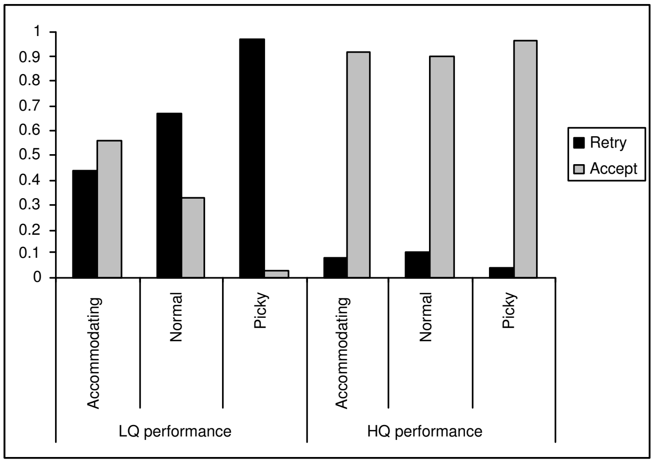


Figure 6.7: Conditional Probability Table(CPT) over user behaviors(discrete) – “Retry” and “Accept”. Each user type is represented numerically with regard to Low Quality(LQ) and High Quality(HQ) system performance(recognition error rate). The Y-axis represents the probability of user behavior conditioned on user type and system performance.

is relatively constant from user to user, the error that users consider acceptable is not, as demonstrated by the variable degree of retries.

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, Sec 6.3.1), we empirically acquired the Conditional Probability Table(CPT) over all the 15 interactions as shown in Figure 6.7. 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), other behaviors ("Accept") dominate covering over 90% in most cases, and allowing us very small amounts of data for observing the "Retry" behavior. However, we can still see that the *Picky* users tend to be more selective than others - they "Retry" less when there is a high quality system performance.

User Behavior Model with the Transonics System

Since in our analysis we observed that the system error alone can not 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 the Table 6.2; 2) The subjects are native speakers(U.S. 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 and adaptation levels are expected to be the same based on the given environment (trained with equal time and materials and provided the same experimental environment for equal time).

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 purpose, the threshold is acquired in a n-fold 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.

6.3.2 A Dynamic Bayesian Network User Behavior Model

A dynamic Bayesian network is a promising representation for modeling the inter-casual relationships of "Retry" behavior with temporal information. The network has been highlighted in the user modeling field across various applications. The Lumiere project [30] utilized Bayesian models for capturing the uncertain relationships between the goals and needs of a user. Conati [12] used Bayesian network to model a student

```
Input: User behavior("Retry" or "Accept") and HQ/LQ recognition information.
Output: The most believable user type
Initial: User types with the same probability
   Step1: The probability of each user type is given by the Bayesian reasoning.
   Step2: Update the prior of each user type
   Step3: Check whether the belief of the highest user type probability is enough
   Step4: If it is not enough to be believed, go to the Step1
Return A user type with the highest probability
```

Table 6.4: 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.

for an automated tutoring system which assesses the knowledge, recognizes plans and predicts actions of each student. Recently, Grawemeyer [28] 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 [23].

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 [30], 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 6.4.

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, 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 [40, 47].

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:

$$\begin{aligned}\sum_{i=1}^{n} P(UT = ut_i) &= 1 \\ \sum_{i=1}^{m} P(B = b_i) &= 1 \\ \sum_{i=1}^{k} P(F = f_i) &= 1 \qquad (6.1)\end{aligned}$$

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).

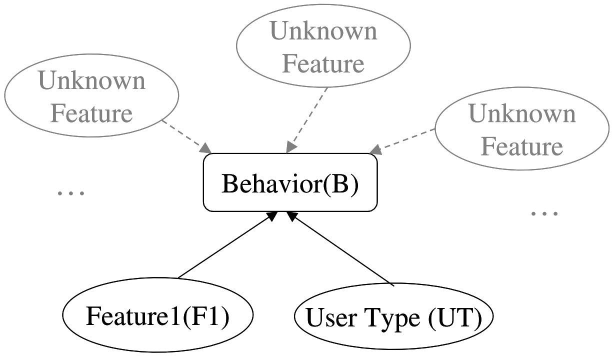


Figure 6.8: A generic directed graphical model; the Bayesian network represents the relation in which a user behavior(B) is influenced by a user type(UT) and a feature (F1). There may be unknown features such as emotions and skill level but only one feature is considered for the suggested model.

The Bayesian network in Figure 6.8 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 different effect on the user behavior. Prior work has demonstrated that fewer features are better for improved accuracy/performance [13], particularly in small data-sets. Also, unimportant features can be eliminated by utilizing probabilistic measures related to the features [91]. 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.

Based on this general procedure, an actual sequence of stepwise conditional probabilities is formed as in the equation (6.2) 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.

$$P(B,UT,F) = P(B|UT)P(UT)P(B|F)P(F)/P(B) \tag{6.2}$$

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 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.

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, by observing a one-time accommodating behavior of a user is not enough to decide a definite user type while the observation of some consistent behavior over time strengthens the belief of the user's type. This idea is formulated as a dynamic Bayesian network (DBN) shown in Figure 6.9. 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.

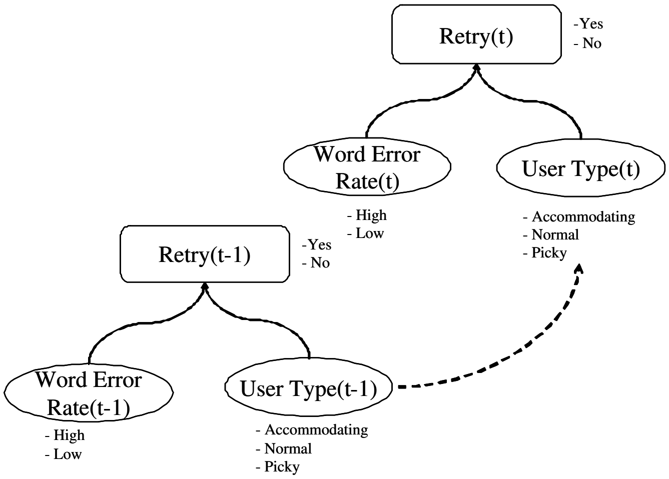


Figure 6.9: A dynamic Bayesian network is used to infer a user type over time in the mediated channel. The belief of a user type becomes strengthened as the interaction progresses.

During training, we employ the complete interaction to reason on the user type by using the Maximum Likelihood Estimate (MLE) as in equation (6.3).

$$P(B|F,UT) = \frac{P(F,UT,B)}{P(F,UT)} \tag{6.3}$$

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). $(P_{\text{WithNormal}})$
- 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:

$$\mu = \lambda \, \frac{(\text{Turn Number})}{(TotalNumberofTurns)} \tag{6.4}$$

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

$$\begin{aligned} P_{\text{SameType}}(n) &= P_{\text{SameType}}(1+\mu) \\ P_{\text{WithNormal}}(n) &= P_{\text{WithNormal}}(1-\frac{1}{3}\mu) \\ P_{\text{Opposite}} &= P_{\text{Opposite}}(1-\frac{2}{3}\mu) \end{aligned} \tag{6.5}$$

	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		

Table 6.5: Values of transition priors. The parametrization allows 4 variables to represent nine time-varying priors, thus allowing estimation from limited data.

Table 6.5 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 Equation (6.6) by applying Bayes' rule.

$$P(UT|B,F) = \eta P(B|UT,F)P(UT) \tag{6.6}$$

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 Equation (6.7), 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.

$$\text{argmax}_i \;\; P(ut_i|B = b_1, F = f_1) = \text{argmax}_i \;\; P(B = b_1|ut_i, F = f_1)P(ut_i) \tag{6.7}$$

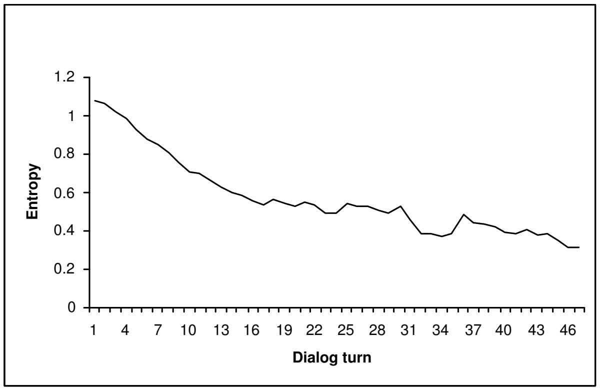


Figure 6.10: Entropy of three user types becomes lower as the dialog turn increases. The threshold of deciding the final user type can be set based on this tendency under a dynamic Bayesian reasoning.

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 *"Threshold"* in confidence. 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 6.3.3 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 [61]. This may be a more objective and concrete measure of convergence and more appropriate for real-time implementations. As in the Figure 6.10, 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 inferrable.

6.3.3 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 13 out of the 15 dialogs 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.

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 Figure 6.11. We can observe that the belief of the *Picky* user type is strengthened over time and is detected early on 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 (Figure 6.11) suspended the flow of conversation in many more cases compared to other users by being very selective.

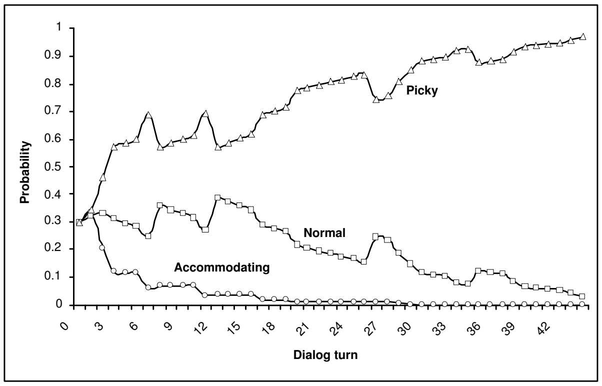


Figure 6.11: The belief that the user type is "*Picky*" is strengthened over time in this example data set.

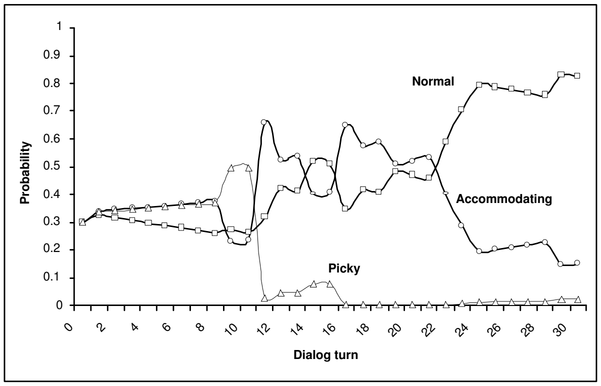


Figure 6.12: The belief that the user type is " *Normal*" is strengthened slowly over time.

Analysis of the *Normal* User Type Inference Result

Figure 6.12 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 boundary of two types, thus causing the DBN to take longer to converge.

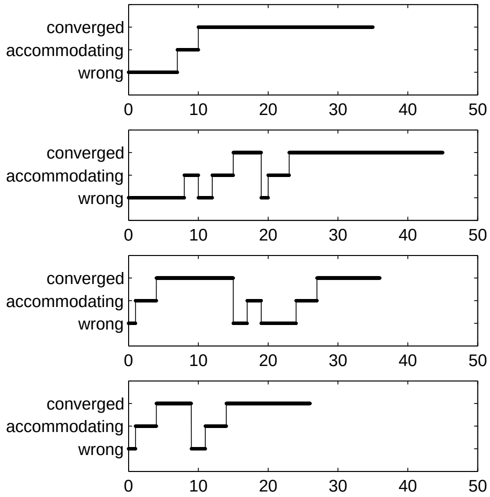
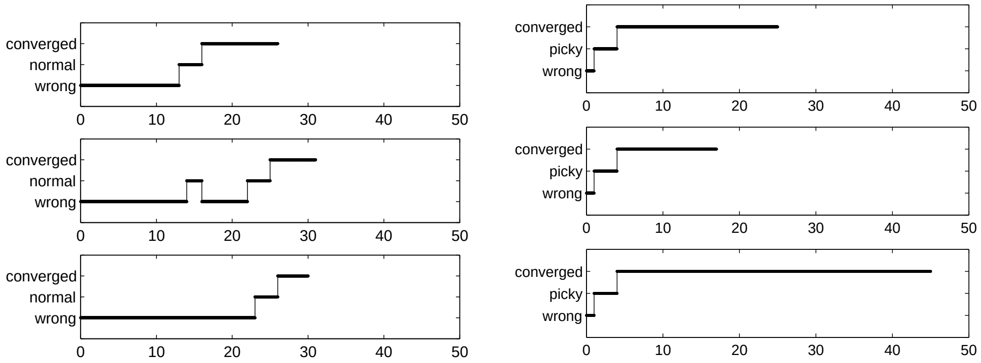


Figure 6.13: Inference on the data of various "Accommodating" user types in the corpus. X-axis indicates the dialog interaction turn. Y-axis indicates three levels of prediction results – wrong, accommodating, and converged to accommodating user types.



(a) Normal

(b) Picky

Figure 6.14: Inference on the data of "Normal" and "Picky" user types over the dialog turns.

Analysis of Successful User Type Inferences

In this subsection, we present the analysis of successful user type classifications by the suggested model(13 out of 15 interactions in our dataset were successful). Figure 6.13

and Figure 6.14(b) represent the identification of the 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 shortly 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 tailing off those of the other user types(Figure 6.14(a)). This implies that it took comparatively more time to be in middle point, in terms of the number of retry/accept under low/high system performance, between the two extremes.

6.4 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, new *T*ransonics system (now, *SpeechLinks*) was used, and English user behaviors were analyzed. The motivation comes from the observation in which users using a mediated device, sometimes, communicate in unnatural fashion: they are extremely picky or accommodating to system errors. Picky type users tend to reject even small recognition errors which do not affect on overall meaning transfer from user-spoken utterance to machine-generated utterance. In the opposite situation, accommodating type users tend to accept even critical recognition errors, which breaks natural conversations between users by causing completely 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.

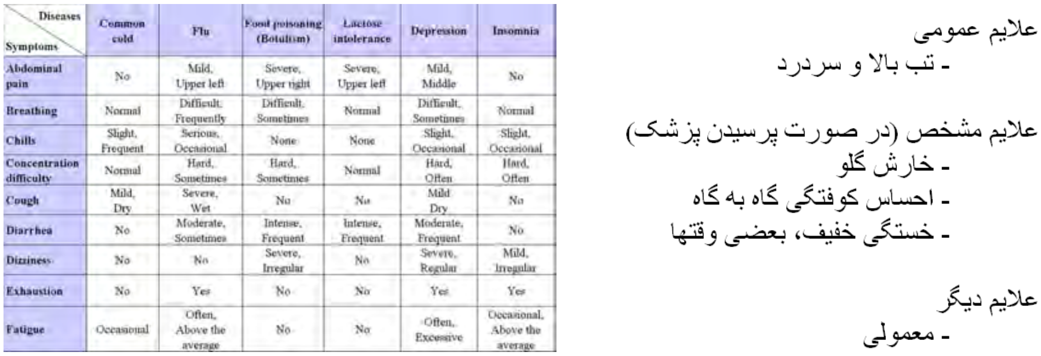
6.4.1 Experimental Setup

Participant and Experimental Domain

To hire native speakers of English, we put a recruiting flyer on campus. Eight English speakers were paid $15 US dollars per hour. They were four males and four females and the age was between 20 and 28. All of them were undergraduate and graduate students at University of Southern California (USC). Two Farsi speakers were contracted students with the *SpeechLinks* project. Farsi speakers were one male and one female with the age of 21 and 24, and undergraduate students.

In total, 32 interaction sessions were collected from 8 native Speakers of English interacted with 2 native speakers of Farsi. For each interaction session, one native speaker of English and one native speaker of Farsi performed a diagnosis of the disease based on the provided scenario. The experimental time of each interaction session was approximately 30 minutes.

The domain of the experiment was doctor's medical diagnosis of the disease of a patient. Native speakers of English played a role of doctor and native Farsi speakers played a role of patient. Before the actual experiment, we gave one hour training session to English speakers and it included how to do a diagnosis of the disease with the supplied materials: the doctor's diagnosis manual table (An example on the left in Figure 6.15) and the instruction of the experiment. For Farsi speakers, we gave enough instructions to use the system and to play a role of patient with the disease symptom card (An example on the right in Figure 6.15). The purpose of the experiment was



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

Figure 6.15: Example materials for the experiment: 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 is presented on the right.

rather to study English speaker behaviors reacting to agent feedback than to study Farsi speaker behaviors. The mission of the English speakers (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 8 English speakers during the experiment.

Scenario

We prepared four scenarios for the experiment using four diseases (flu, SARS, depression and hypertension), and each experiment team (one native speaker of English and one native speaker of Farsi) used these four scenarios in the same order during the experiment. For each scenario, we provided a doctor's diagnosis manual table consisting of twelve (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 as realistic as possible using the medical diagnosis information from **“http://www.medicinenet.com.”**

Farsi speakers (the patient role) were given a symptom card which provided only a few symptoms of the disease. On the right image in Figure 6.15, 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 reason about a disease of the symptom card of a Farsi speaker.

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

Procedure of Experiment

The experimental procedure was designed with two tasks, borrowing the idea of the evaluation method in the user modeling work by [47]. Figure 6.16 shows this experimental procedure. In “Task A”, native speakers of English performed the interaction session of “without feedback” first and the session of “with feedback” later. In “Task B”, native speakers of English performed the interaction sessions in the reverse direction. In each task, 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). For more objective data collection, we assigned Farsi speakers evenly to the two tasks: each Farsi speaker participated in “Task A” 4 times, and the “Task B” 4 times. In total, we collected 32 interaction sessions from this experiment.

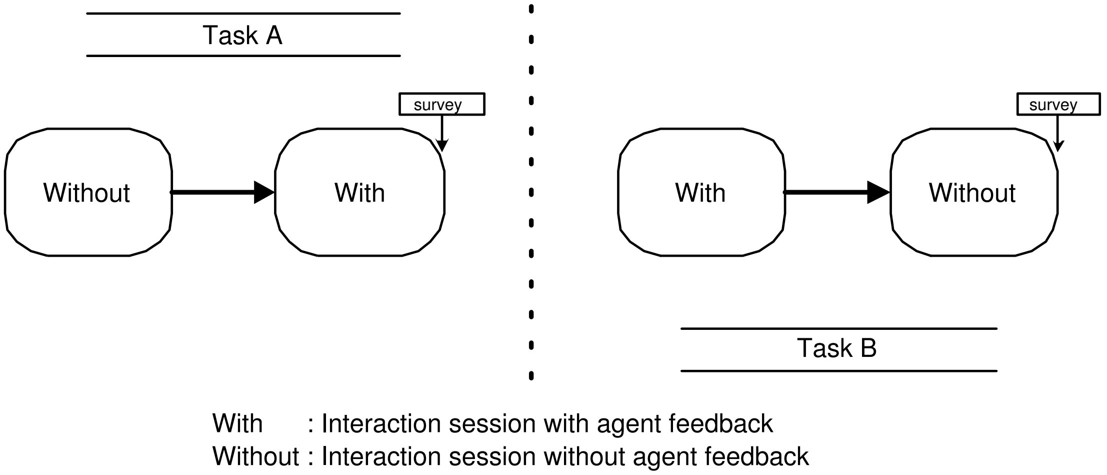

Task A
survey
Without
With
With
Without
survey
Task B
With : Interaction session with agent feedback
Without : Interaction session without agent feedback

Figure 6.16: All 8 English speakers performed both “Task A” and “Task B” with 2 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.

For evaluation purpose, we collected 5 survey questionnaires from each participant during the experiment. One is the initial survey about demographic information of the participant and user perception on many subjects, such as user type and error tolerance level and past speech interface experience. After each interaction session, an questionnaire was given to each participant for the evaluation of system performance in multiple dimensions, such as user satisfaction and interaction efficiency. In total, 4 evaluation questionnaires were collected from each participant. Detail analyses of questionnaires are provided in the section 6.4.2.

Each session lasted for thirty minutes approximately - we gave a notice to them (to finish the session) when they did not finish the session in thirty minutes. After finishing two sessions (with feedback and without feedback in the order or in the reverse order), participants gave us opinions about the experiment.

	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 re-trying"	"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."

Table 6.6: 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.

All the interaction sessions were recorded with video cameras. 'Sony Hi-8' and 'Sony high definition' video cameras were used for this purpose. We analyzed thirty two (32) interaction sessions in the video data in terms of identifying user types with their behaviors and, user behavior changes and system performance.

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 detects one of the two user types with high probability, it triggers the corresponding wording of agent feedback as in Table 6.6. 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 accommodating or picky user type 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 6.3.2. At each turn in interaction sessions, 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.

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 6.3.2. The correlation between ASR confidence level and WER was mentioned and studied in [25, 104]. ASR confidence level was computed using features at multiple levels, such as weighted acoustic model and language model scores.

6.4.2 Experimental Result

We present the results of 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.

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 wend wrong and it helped for smooth conversation flows and information delivery. In particular, the participants commented that agent feedback helped to get less frustrations caused by repetitive errors. One of 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.”

Subjective Measure 2: Video Analysis

By analyzing the video data of 32 interaction sessions, we subjectively identified user types of 8 English participants: 7 participants were picky and 1 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 comparison between the classified users by the analysis of video data, and those by the analysis of log data is presented in Table **??**.

The analysis of video data suggested us a trend of user accommodating 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 agent feedback. We conjecture that they already knew what the agent feedbacks were and perceived 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.

Subjective Measure 3: Questionnaire Analysis

We collected five questionnaires from each participant and the Likert-scale questions were given to the participant. 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 6.7).

One finding from the initial questionnaire is that some users did not have speech interface experience at all but others had already enough experience. To reduce this gap, we gave a one hour training session to all participants, which included how to use the

Likert-scale questions	mean	std. dev.
Speech interface experience(0:none - 10: more than ten 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

Table 6.7: The statistics collected from the Likert-scale questions of the initial survey given to the participants. We measured users' own perception about their ability of dealing with general technology and speech interface, utterance length, and error tolerance level.

system. Another interesting finding was that the error tolerance level in the communication with human was higher than that with computers, indicating that they are more generous in tolerating errors in the communication with human than with computers.

In the other four questionnaires, we measured (after interaction sessions) 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 (8.71(1.3)) and easy to learn (8.18(1.1)) but not foolproof (3.5(1.0)).

To measure the effect of agent feedback, the comparison of user satisfaction between the interaction session with agent feedback and the interaction session without agent

	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

Table 6.8: Overall user satisfaction (Likert scale – 1: worst 10: best) after interaction session in each of the two tasks (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)).

feedback is presented in Table 6.8. In addition, this comparison was conducted separately in each of the two tasks. Higher user satisfaction was observed in the interaction session with agent feedback across the two tasks. More specifically, to find out statistical significance, “Paired Sample T-test” was performed on each Task and we acquired p values, 0.264 from “Task A”, and 0.041 from the “Task B”. The observed significance level of the “Task B” is enough to say the statistical difference between two interaction sessions (p ¡ 0.05).

Fundamental statistics collected from the questionnaires which support for the results of Table 6.8 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 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).

Objective Measure: Log Data Analysis

In this section, we investigated user behaviors accommodating to errors, and effects of agent feedback on the interaction efficiency. For measuring the interaction efficiency, we attempted to observe some objective metrics which were introduced in the PARADISE framework [99]. The metrics were to measure user satisfaction, and the efficiency and cost of interaction sessions of spoken dialogs. However, most of the metrics were not applied to our data: we conjecture the reason that first, our system is a mediated device between two interlocutors – but, the PARADISE framework is for evaluating human-machine spoken dialogs, and second, the domain of our system is medical diagnosis which does not require exact parameters in the metrics of the PARADISE framework.

Before going into detail, it may be interesting to know 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 minute and 36 seconds (3 minute and 2 seconds) with agent feedback, and 32 minute and 27 seconds (4 minute and 13 seconds) 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 triggering agent feedback in an interaction session was 10.7 (7.87) – excluding the interaction sessions without agent feedback.

In the video analysis results, we observed that only one participant was the accommodating type, who endured relatively more recognition errors compared to the other 7 participants. In the log data of 8 participants, we measured retry rates of the participants under low system performance, and the same participant was classified as the accommodating type by the *k-means* algorithm: we only classified the participants into two

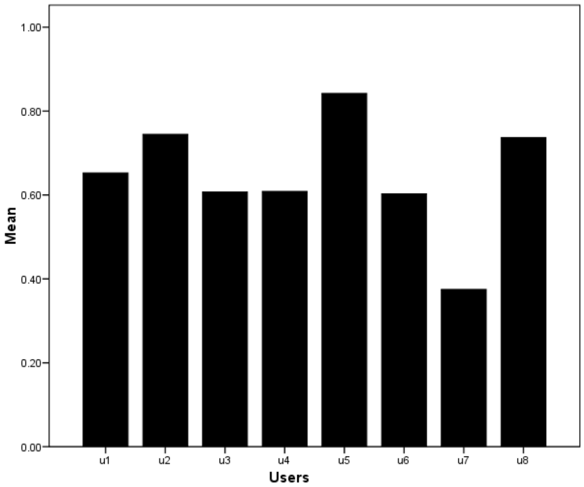


Figure 6.17: User retry rates over the interaction sessions when the ASR performance is low. Interaction sessions without agent feedback were investigated. Seven users was clustered as picky and one as accommodating.

types, accommodating and picky. The low system performance is the low ASR confidence level, and we investigated the interaction sessions without agent feedback for this analysis. The user retry rates over the interaction sessions are presented in Figure 6.17.

Interesting part in the analysis of log data was whether we could find interaction efficiency (e.g., naturalness of interactions or smooth conversation flows) in the sessions with agent feedback. We defined this interaction efficiency as the number of normal user type appeared during an interaction session: the more normal type appears during the interaction session, the more efficient the interactions are. The reason for this definition is as follows. Normal type users are not extreme to accept/reject system errors so we expect to avoid extreme cases (such as severe repetitions) in their interaction sessions. Intuitively thinking, we have smooth conversations with people when we are in normal type. In one of the analysis results, the normal user type appeared more during the interaction sessions with agent feedback than during the interaction sessions without agent feedback as in Table 6.9.

without agent feedback	with agent feedback
0.37 (0.14)	0.44 (0.14)

Table 6.9: Percentage (with standard deviation in parenthesis) of normal user type appeared during the two interaction sessions: with/without agent feedback. More normal user type during the interaction sessions indicates more efficient interactions.

Another interesting aspect to study is to investigate the effect of agent feedback on user behaviors in better direction, and it contributes to the efficient 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 6.10). 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 user behavior whether it was changed from the previous turn or not. Note that it is a possible chain of errors when the two extreme user types were triggered by the dynamic *Bayesian* reasoning. As shown in Table 6.10, 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.

without agent feedback	with agent feedback
0.31 (0.21)	0.40 (0.16)

Table 6.10: Percentages of user behavioral change from the previous turn in the 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.

6.5 The Interpersonal Channel

6.5.1 Early Attempts at Analyzing Speech Accommodation

Human-Human conversation research for spoken dialog systems relies on multidimensional analysis of a variety of speech and language characteristics. Shriberg [94] describes 4 properties – *punctuation, disfluencies, turn taking and hearing speaker's emotion/state* based on human-human spontaneous speech conversation analysis. In addition to the features directly stemming from the interaction, many other user dispositions, such as mood (longer term emotions), culture, gestures and eye-gaze [83], are considered as appropriate features for user modeling issues.

To investigate user behavior in the interpersonal channel, we base our work on Speech Accommodation Theory (SAT). SAT provides insights into how a human modifies his/her speech style when interacting with another human (agent) and shows convergence and divergence of verbal and non-verbal behavior [26]. Speech accommodation is regarded as a good skill for acquiring social approval or acceptance – and is likely to correlate with high user satisfaction in the system evaluation. For instance, prior research has found that speakers tend to show lexical adaptations(coincidental overlap) in conversation in mediated contexts and suggest that users' lexical accommodation can be utilized to improve system performance [20]. Matessa [62] showed that accommodating speech style boosts the communication efficiency of the conversation mechanism as well as helps in social approval creating greater rapport.

A challenge in the present domain is the concurrent face-to-face (interpersonal) communication channel being open between the interlocutors. The users can exchange explicit non-verbal gestures, or implicitly modify their verbal behavior such as to accommodate each other's speech. Despite the limited amounts of available data, we have attempted to analyze the interpersonal channel of this translation system.

We first analyzed the 15 simulated doctor and patient iterations in an attempt to identify *prosodic accommodation*, but the data did not support this hypothesis. We believe this can be attributed to the effects of the mediation channel that regenerates the prosody (with synthetic speech) and removes some of the social aspects of the interaction, thus making conclusions from such a limited data set impossible. Another point of view on this is that it is artificially generated data. These are role playing doctor-patient interactions, even though we tried to make the experiment as real as possible by hiring medical professionals and standardized patient actors.

We then performed an analysis testing the participants *utterance length* accommodation. This also did not give us reliable results to assert the hypothesis of human-human utterance length accommodation but nevertheless provided us with information regarding user accommodation to problematic conditions as discussed in the following section.

6.6 Discussion

6.6.1 Performance Dependency between Retrying and Utterance Length

In the data set, we observed how users differ in expressing themselves and how that is reflected in the overall utterance length. The average length of doctor utterances is 4.45 with the Standard Deviation(STD) 0.39 and that of patient is 2.42 with STD deviation 0.62. Doctors mostly ask direct questions such as, “Do you have difficulty breathing?” and patients usually keep their answers short. The length of doctor utterances is more consistent than that of patients because, they were trained to ask specific questions relevant to their field and were trained to use the translation system, which handles single concept utterances best. This is a barrier in revealing the user type of the doctor based on utterance length. By contrast, a patient is less constrained and free in revealing his/her personal characteristics. Longer utterances will implicitly include details or extra information.

Human behavior is an extremely complex process dependent upon a multitude of variables. The two specific user behaviors (*Retry* and *Speech Accommodation*) studied above can be further analyzed in terms of their correlation.

It is observed that users tend to employ shorter utterances following errors in the mediation channel. On average, the utterance length decreased by 60% following a “Retry” but increased by 62% following a successful exchange.

Utterance length also affects recognition performance since longer utterances tend to generate higher error rates. An utterance length increase was followed by a significant probability (63%) of “Retry”. A case can be made on the mutual dependence of the mediated and interpersonal channels based on these observations.

In summary, based on the observations above, we concluded that it is difficult to draw trustworthy speech style accommodation results, especially regarding the direct, human-human communication path. However we believe that there is speech accommodation taking place in the human-machine interaction channels. This is due both to the extra effort required by the users in accommodating the device and the current implementation of cross-lingual mediation devices; people tend to put emphasis on how to operate the device, thus affecting their communication styles. Additionally, since the conversation is task oriented, users are less likely to try to overcome the mediation channel barriers in order to exhibit socially acceptable behavior, such as that described by the SAT.

6.6.2 Data Size and Data-Driven Modeling

One of the major challenges of an empirically-based user modeling study is the availability of data. It is especially important to note that it requires a huge effort to collect, process and interpret the complex data from these bilingual spoken interactions. 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 test set [**?**]. 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.

6.6.3 Lessons from the Online Experiment with Agent Feedback

In designing a mediated device, it is important to have a good understanding of user model, thus be able to appropriately modify the communication strategies, for example, by taking system initiative. These system initiatives must be well founded on robust user models to ensure minimal user disruption. We designed triggering agent feedback in this fashion (not disruptive). However, some participants in the online experiment using agent feedback commented that they needed the feedbacks mostly in the early time of interaction sessions and the repetitive feedbacks might 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 offline experiment can be substituted by ASR confidence level. This assumption is considered acceptable widely in speech technology community. However, it is still debatable whether, in what situation, with what features, we can accept this assumption.

6.6.4 Future Directions

Determining the subject of accommodation should also be considered in a design phase. In our case, a doctor is more likely to accommodate the patient's speech style because he/she provides a medical service to a patient [87, 29]. In addition, as we observe, there is another layer of accommodation stemming from the user's accommodation towards the device.

In addition to utterance length, there are many features that can be considered for analyzing accommodation behaviors including speech rate, pausing, culture, social status, etc. It is an interesting topic to consider the information content of such features in future research.

6.7 Conclusions

The paper addressed user behavior modeling approaches in a machine-mediated setting involving bi-directional 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. For realistic application of the model, online experiment with agent feedback was performed and results with subjective and objective measures were reported.

We modeled a specific user behavior with 3 user types, *Accommodating*, *Normal* and *Picky*. The granularity of user type can be adjusted according to the desired response. 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 the users are extremely tolerant of errors. In the offline data, we showed that one of 3 types becomes obvious as a user keeps his/her consistent behavior under the same condition belonging to a specific type. This model can be utilized for the design of an efficient error handling mechanism; in previous research [82], a correct interpretation of user's goal (intention) was helpful in dealing with errors in human robot dialogs. Ultimately, we believe 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 [99]. 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.

We presented ideas for future work, including a first attempt at addressing speech accommodation issues, which, however, are currently inconclusive due to limited data availability. We notice, however, in analyzing utterance length statistics that there is an apparent correlation of the retry behavior with utterance length changes. This, we believe, suggests an accommodation response to the device limitations, but one that needs further investigation. As part of on-going work, we described a preliminary design of a conversation assistant agent based on the suggested user model between two interlocutors. Evaluation on user satisfaction, conversation efficiency and system performance improvement will be the next steps.

Chapter 7

Conclusion and Future Work

“Collection of the information about users is the key for better system performance and higher user satisfaction.”

The study presented four user modeling work for human-machine spoken interaction and mediation systems (speech-enabled systems), especially focused on user models in error conditions. Under error conditions of the speech-enabled systems, the systems, not just wait for users to correct the errors, need to actively take initiative to help users to get over the errors. For analyzing and modeling work, We utilized the data logged by two speech-enabled systems: (1) *DARPA communicator* which is a human-machine Spoken Dialog System (SDS), and (2) *Transonics (new name: SpeechLinks)* which is a translation driven human-human Spoken Mediation System (SMS). The domains of these systems are travel agency service(the system 1) and medical diagnosis(the system 2) respectively. The research investigated the logged data of users using the speech-enabled system, and modeled users based on the statistics. For evaluation purposes, user survey questionnaires and user interview data as well as statistical analysis of log data, were used in the study. The goal of the study is to contribute to the implementation of speech-enabled systems which can handle diverse users under error conditions. Eventually, the user modeling work of this dissertation is expected to lead the higher user satisfaction and enhanced system functionalities that provide user-centered functionalities and services.

User modeling work started with the analysis of user behaviors under error conditions of a spoken dialog system – when things do not go well in the communication chain. In particular, we examined categories of error perception, user behavior under error, effect of user strategies on error recovery, and the role of user initiative in error situations. A conditional probability model smoothed by weighted ASR error rate was proposed. From the analysis we found: (1) users discovering errors through implicit confirmations were less likely to get back on track (or succeed), and took a longer time in doing so than other forms of error discovery such as system reject and reprompts; (2) Further successful user error-recovery strategies included more rephrasing, less contradicting, and a tendency to terminate error episodes (cancel and startover) than to attempt at repairing a chain of errors. These analyzed results of user behaviors can be utilized for the better design of an efficient error handling mechanism in building a spoken dialog system.

One of major sources of errors is due to incorrect automatic speech to text conversion. In interactive application, it is important that these errors do not impact the overall concept transfer between users and the system. For speech-to-speech translation system, this becomes more important because the evaluation metric of the translation relies mostly on the meaning (the number of concepts) delivered through the information exchange channels. In this regard, user behaviors were analyzed in terms of the number of concepts transferred through the mediating device, *Transonics*. To clarify the definition of how many concepts are transferred in the utterances produced by the system (from original user utterances), the *Concept Matching Score (CMS)* was proposed, and was defined based on “adequacy” levels which asses the quality of translations by the system (proposed by Linguistic Data Consortium (LDC)). The results showed that while some users require perfect representation of what they said in order to allow transfer, others accept degradation to some extent. An appropriate system strategy is required to recognize this behavior and guide users towards optimum performance points. In addition, we compared machine translation performances between the unimodal (speech) interface setting and the multimodal (speech and visuals) interface setting by measuring the CMS. The analysis showed that employing multimodal interface improves translation quality by 24%.

In a viewpoint of a system designer, users are considered as a changing subject over time while their using the speech-enabled systems. They change behaviors, intentions, or even goals while using the speech-enabled systems. In this regard, we reported the analysis results of users who used a multimodal interface of a speech-to-speech translation system during the 4 weeks. Three sets of collected data were investigated for the analysis purpose: user interview data, user survey questionnaire, and log data of the system. By using the user interview data and the user survey questionnaire, we reported: (1) users incorporated the strategies to cope with system errors in unsuccessful turns, such as repeat, rephrase, change topic, and start over; (2) users perceived their proficiency in using and learning the system improved during the first three weeks. For the analysis of log data of the translation system, meaning of utterance was considered important in the study. We used the Concept Matching Score (CMS) as introduced in the previous chapter. Using this metric, we first reported the distribution of transferred meaning level in two cases; successful and unsuccessful interaction turns of conversations. 91% of utterances in the successful turns contained more than half the meaning of the original user speech, and 90% of utterances in the unsuccessful turns contained less than/equal to half the meaning of the original user speech in our data. Second, we investigated the meaning transfer level by the multimodal interface comparing with that by the speech-only interface. We observed improvement of meaning transfer by 33% and by 11% through the multimodal interface in comparison with two speech-only interface settings respectively; one without and the other with filtering unsuccessful interaction turn. Third, we reported that users gradually accepted machine-produced utterances more during the 4 weeks. Further analysis showed that users became more accommodating to the system errors after having experiences of using the system, such as functional word insertion errors which usually does not impact on the final translation quality. In general, users of speech-enabled interface have a strategy to deal with system errors, which tends to change the length of speech. In the report, the length of user speech increased after successful interaction turn, and decreased after unsuccessful turn. During the 4 weeks, average length of user speech was reduced gradually in the later 3 weeks.

From the analysis results above, we observed that users differ in accepting system errors – the system errors are word-level lexical errors which can be identified easily by the system. Some people are more accommodating to the system errors than others. In this regard, we defined and clustered users as *Accommodating*, *Normal*, and *Picky* user types with features, the word error rate and the user retry behavior. The user models were defined and evaluated in the setting of speech-enabled mediating device, *Transonics* (now, *SpeechLinks*). A dynamic Bayesian network was proposed to model three user types with two features (as described above), which is an inference mechanism to automatically determine the user type in offline and online experiments. To validate the model, we performed both offline and online experiments using the log data of *Transonics* and the agent feedback implemented in the *SpeechLinks* system. The experimental results using offline data showed that one of the 3 user types 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 analyzed recorded video, user interview data, survey questionnaire, and the system log data of this online experiment. The analysis results showed high user satisfaction and enhanced interaction efficiency for both users and the system.

One of the challenging issues in the study of this dissertation was communication between two persons. The case of mediated interactions is more complex than a human-machine interaction setting. Under a human-human interaction with a mediating device setting, there are two channels delivering the information: (1) the mediating channel, and (2) the interpersonal channel. We attempted to address modeling users interpersonal adaptation in the "interpersonal" channel, focusing on the design of a model based on the users' utterance length. This analysis implies that the average utterance length of a user reflects a specific user type and can be used for dealing with interpersonal adaptation. The *Speech Accommodation Theory* [27] was adopted to support the argument that greater degree of utterance length accommodation is related to the higher user satisfaction in the Human-Human interaction. Some results were presented in the analysis part; user utterance length was related to the recognition errors and users did not show perfect interpersonal adaptation because of this recognition error effect. In the future, more investigations on correlations and effects between the "mediating" channel and the "interpersonal" channel are demanded.

Because of the error-prone property of statistical processing in speech technology, speech-only interface may not enough to convey information. Multiple user input usages will be desirable in this case for practical applications using speech technology. This multimodal interface would provide alternative options to users who may have problems in using speech technologies. In this regard, we would like to extend our work to investigate multimodal user behaviors under a setting of two person's communication (who are speak different languages) using a speech-to-speech translation system. In particular, it will be challenging to study the aspects of two person's cooperations using a multimodal interface, comparing those between two settings, a machine-mediated communication (translation device) and a human-mediated communication (human translator).

Appendix A

Tag Set and Guidelines

Each category of tag is preceded by the conventions for applying it. Examples of utterances that would receive each tag are provided in the tables where practical.

A.1 System Tags

Definition: Clues by which the user becomes aware of an error.

Put tag on system said: line inside and outside of error segments wherever the phenomena occurs. Multiple tags are okay.

A.2 User Tags

Definition: Users response to errors.

Put tag on “user said” lines inside and outside of error segments wherever the phenomena occurs. Multiple tags are okay.

A.3 Task Tags

Definition: Tags about the state of system/user interaction.

expl :	Explicit confirmation. User is asked to confirm certain input. An example; Was the arrival city Wichita or London ?
impl :	Implicit confirmation. The system repeats the users last input to introduce the following prompt. An example; A flight from Miami. Where do you want to go?
reject :	Rejection. The system tells the user that the recognizer either did not hear or did not understand the last input. An example; Sorry, I misunderstood. Please say the name of the city or airport you wish to depart from. (reject + aid)
aid :	Aid. The system instructs the user to give the input in a certain way. Often used with reject and also often used in the first system prompt. An example: Try asking for flights between two major cities.
nonseq :	Non Sequitur. An inappropriate system response gives the user evidence of error. An example; User said: [throat_clearing] I would like to book a flight from Columbus Ohio to Phoenix Arizona to arrive before six p. m. on October fifth System said: Where are you departing from? (nonseq)
sysrepeat :	System Repeat. The system repeats its last prompt. (a pathological example where ASR looks ok, but response is not accepted) An example; System said: What is your destination? User said: Phoenix Arizona Recognizer heard: Phoenix Arizona System said: What is your destination? User said: the destination is Phoenix Arizona (user rephrase) Recognizer heard: The destination is Phoenix Arizona System said: What is your destination ?

Table A.1: Clues by which the user becomes aware of an error.

repeat :	User repeats exactly what they said in the previous turn.
rephrase :	User rephrases the last input, modifying choice of words, their order, etc. An example; System said: What time would you like to depart? User said: early Recognizer heard: early (err) System said: Sorry, I misunderstood. Please give the approximate time you would like to depart. (reject + aid) User said: eight a. m. (rephrase) Recognizer heard: eight a m
contradict :	The user contradicts the system, often as a barge-in. An example; System said: What time do you want to leave phoenix (impl) User said: no I don't want to leave Phoenix I'm starting from Columbus Ohio
frust :	The user shows signs of anger (frustration), contempt, disgust, and frustration. An example; Oh my god [uh] can we start over (frust + startover)
chngreq :	The user tries different dates, different cities in the same state/country in an attempt to circumvent an error. An example; System said: Flying from Dulles. What city are you flying to? User said: Hilton Head South Carolina System said: At the Hilton. What city are you flying to? (nonseq + sysrepeat) User said: Hilton Head South Carolina System said: At the Hilton. What city are you flying to? (sysrepeat) User said: Savannah Georgia (chngreq)
startover :	The user has the system start over from scratch using the start over command used by most of the systems. An example; System said: Sorry, I didn't understand that. What city are you flying to? (reject + sysrepeat) User said: START OVER

scratch :	Cancel the last input. Recognizer heard: I want to fly from Seattle to San Diego (err) User said: I want to fly from Seattle to Tokyo System said: flying to San Diego on what date? (impl) Recognizer heard: scratch User said: scratch (scratch) System said: Canceling only this flight segment request. Please give me the details of the request for this segment again. For example, you can say, I want to fly from Boston to Dallas next Sunday morning. (aid)
ask :	The user directs a question to the system or asks for help. An example; Do I have to fly to Rome to get to Berlin?
acq :	The user continues a dialogue without trying (acquiescence) to correct errors. May end an error segment without getting back-on-track (in this case the tag may not be on a user said line). When there is an acq, a note is added about what was acquiesced to.
hangup :	The user hangs up in response to an error (if it is the computer that hung up, place tag on the "system said" line

Table A.2: Users response to errors.

err :	Placed at the “Recognizer heard:” turn (error) where the initial error occurred. Ignore any problem in the requests for ID numbers at the beginning of dialogues. When the error is minor (e.g. late afternoon instead of early afternoon) and the user doesnt try to correct it, use acq instead of err.
bot :	Back-on-track. The user and system successfully negotiated the correction of an error. Placed on the system said line that provides the user (and tagger) with evidence of being totally back on track (at the end of an error segment?never in nested errors).
succ :	The user got the tickets s/he wanted (success) If there are small errors like flight time, then use with acq.

Table A.3: Tags about the state of system/user interaction.

Reference List

[1] J. Aberdeen, C. Doran, L. Damianos, S. Bayer, and L. Hirschman. Finding errors automatically in semantically tagged dialogues, 2001.

[2] J. Allen and M. Core. Draft of damsl: Dialog act markup in several layers. http://www.cs. rochester.edu/research/cisd/resources/damsl/, 1997.

[3] J. Allen, B. W. Miller, E. K. Ringger, and T. Sikorski. A robust system for natural spoken dialogue. In Arivind Joshi and Martha Palmer, editors, *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 62–70, San Francisco, 1996. Morgan Kaufmann Publishers.

[4] J. Allen and C. Perrault. Analyzing intention in utterances. *Artificial Intelligence*, 15(3):143–178, 1980.

[5] A. Batliner, K. Fischer, and et al. How to find trouble in communication. *Speech Communication*, 40(1-2):117–143, 2003.

[6] G. W. Bauer. Interface for user/agent interaction. In *U.S. Patent 5877759, Mar 2*, 1999.

[7] A. W. Black, R. D. Brown, R. Frederking, R. Singh, J. Moody, and E. Steinbrecher. TONGUES: rapid development of a speech-to-speech translation system. In *Proc. of HLT-2002*, March 2002.

[8] T. Bub and J. Schwinn. VERBMOBIL: The evolution of a complex large speech-to-speech translation system. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 1996.

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

[10] J. Chu-Carroll and M. K. Brown. Tracking initiative in collaborative dialogue interactions. In *Meeting of the Association for Computational Linguistics*, pages 262–270, 1997.

[11] J. Cohen. From meaning to meaning: the influence of translation techniques on non-english focus group research. *Qualitative health research*, 11(4):568–579, 2001.

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

[13] M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis 1(3)*, 1997.

[14] L. Deng, Y. Wang, K. Wang, A. Acero, H. Hon, J. Droppo, C. Boulis, M. Mahajan, and X.D. Huang. Speech and language processing for multimodal human-computer interaction. *The Journal of VLSI Signal Processing*, 36(2-3):161–187, 2004.

[15] G. Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of ARPA Workshop on Human Language Technology*, 2002.

[16] L. Dybkjr, N. O. Bernsen, and W. Minker. Evaluation and usability of multi-modal spoken language dialogue systems. *Speech Communication*, 43(1-2):33–54, 2004.

[17] W. Eckert, E. Levin, and R. Pieraccini. User modeling for spoken dialogue system evaluation. In *ASRU'97*, 1997.

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

[19] G. D. Fabbrizio, D. Dutton, N. Gupta, B. Hollister, M. Rahim, G. Riccardi, R. Schapire, and J. Schroeter.

[20] L. Fais. Lexical accommodation in machine-mediated interactions. In *16th COLING*, 1996.

[21] T. W. Finin. Gums: A general user modeling shell. *In A. Kobsa and W. Wahlster editors, User Models in Dialog Systems*, page 411.430, 1989.

[22] G. Foster, P. Langlais, and G. Lapalme. User-friendly text prediction for translators. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, 2002.

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

[24] Y. Gao, L. Gu, B. Zhou, R. Sarikaya, M. Afify, H.K. Kuo, and et al. IBM MASTOR SYSTEM: Multilingual automatic speech-to-speech translator. In *Proceedings of the First International Workshop on Medical Speech Translation, in conjunction with NAACL/HLT 2006*, 2006.

[25] K. Georgila, J. Henderson, and O. Lemon. Learning user simulations for information state update dialogue systems. In *Eurospeech*, 2005.

[26] H. Giles and N. Coupland J. Coupland. *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press, Berlin, Heidelberg, New York, Barcelona, Hong Kong, London, Milan, Paris, Singapore, Tokyo, 2000.

[27] H. Giles, A. Mulac, J. Bradac, and P. Johnson. Speech accommodation theory: The first decade and beyond, 1987.

[28] B. Grawemeyer and R. Cox. A bayesian approach to modelling users' information display preferences. In *User Modeling 2005*, 2005.

[29] J. A. Hall, D. L. Roter, and N. R. Katz. Meta-analysis of correlates of provider behavior in medical encounters. *Med. Care*, 26(7):657–75, 1988.

[30] E. Horvitz, , J. Breese, and et al. The lumiere project: Bayesian user modeling for inferring the goals and needs of software users. In *Proc. of the 14th Conf. on Uncertainty in Artificial Intelligence*, 1998.

[31] E. Horvitz. Principles of mixed-initiative user interfaces. In *CHI*, pages 159–166, 1999.

[32] E. Horvitz and T. Paek. Harnessing models of users' goals to mediate clarification dialog in spoken language systems. In *Proc. of the 8th Int. Conf. on User Modeling*, 2001.

[33] E. Horvitz and T. Paek. Harnessing models of users' goals to mediate clarification dialog in spoken language systems. In *Proceedings of 8th International Conference, UM 2001*, 2001.

[34] K. Isbister, H. Nakanishi, T. Ishida, and C. Nass. Helper agent: designing an assistant for human-human interaction in a virtual meeting space. In *CHI*, pages 57–64, 2000.

[35] D. Jannach and G. Kreutler. Personalized user preference elicitation for e-services. In *Proceedings of the 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'05) on e-Technology, e-Commerce and e-Service*, 2005.

[36] K. Jokinen and K. Kanto. User expertise modelling and adaptativity in a speech-based e-mail system. In *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics ACL-04*, 2004.

[37] A. Bernstein K. Reinecke. Culturally adaptive software: Moving beyond internationalization. In *Proceedings of the HCI International 2007*, 2007.

[38] C. Kamm, D. Litman, and M. Walker. From novice to expert: The effect of tutorials on user expertise with spoken dialogue systems, 1998.

[39] C. Kamm and M. Walker. Design and evaluation of spoken dialog systems. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1997.

[40] C. A. Kamm, D. J. Litman, and M. A. Walker. From novice to expert: The effect of tutorials on user expertise with spoken dialog systems. In *Proc. of ICSLP*, 1998.

[41] J. Kay. The um toolkit for reusable, long term user models. *User Modeling and User-Adapted Interaction*, 4(3):149–196, 1995.

[42] K. Knight and D. Marcu. Machine translation in the year 2004. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

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

[44] A. Kobsa. Supporting user interfaces for all through user modeling. In *Proceedings of the HCI International*, 1995.

[45] A. Kobsa. Generic user modeling systems. *User Modeling and User-Adapted Interaction*, 11(1-2):49–63, 2001.

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

[47] K. Komatani, S. Ueno, and et al. Flexible guidance generation using user model in spoken dialogue systems. *41st Annual Meeting of the ACL 2003*, pages 256–263, 2003.

[48] K. Komatani, S. Ueno, T. Kawahara, and H. G. Okuno. Flexible guidance generation using user model in spoken dialogue systems. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics(ACL2003)*, 1:256–263, 2003.

[49] B. Krulwich. Lifestyle finder: Intelligent user profiling using large-scale demographic data. *AI Magazine*, 18(2):37–45, 1997.

[50] A. Kuenzer, C. Schlick, and et al. An empirical study of dynamic bayesian networks for user modeling. In *Proceedings of the UM 2001 Workshop on Machine Learning for User Modeling*, 2001.

[51] I. Langkilde, M. Walker, J. Wright, A. Gorin, and D. Litman. Automatic prediction of problematic human-computer dialogues, 1999.

[52] M. L. Larson. *Meaning-Based Translation:A guide to cross-language equivalence(Second Edition)*. University Press of America, 1997.

[53] 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 att-darpa communicator mixed-initiative spoken dialog system, 2000.

[54] E. Levin and R. Pieraccini. A stochastic model of computer-human interaction for learning dialogue strategies. In *Proc. Eurospeech '97*, pages 1883–1886, Rhodes, Greece, 1997.

[55] E. Levin, R. Pieraccini, and W. Eckert. A stochastic model of human-machine interaction for learning dialogstrategies. *Speech and Audio Processing*, 8(1):11–23, 2000.

[56] G. A. Levow. Characterizing and recognizing spoken corrections in human-computer dialogue. In *COLING-ACL*, pages 736–742, 1998.

[57] F. Linton and H. P. Schaefer. Recommender systems for learning: Building user and expert models through long-term observation of application use. *Journal of User Modeling and User-Adapted Interaction*, 10(2-3):181–208, 2000.

[58] D. Litman and S. Pan. Empirically evaluating an adaptable spoken dialog system. In *Proc. of UM'99*, 1999.

[59] J. Boland E. Bratt J. Garafolo L. Hirschman A. Le S. Lee S. Narayanan K. Papineni B. Pellom J. Polifroni A. Potamianos P. Prabhu A. Rudnicky G. Sanders S. Seneff D. Stallard M. Walker, J. Aberdeen and S. Whittaker. Darpa communicator dialog travel planning systems: The june 2000 data collection. In Proc. Eurospeech, 2001.

[60] X. Ma and C. Cieri. Corpus support for machine translation at LDC. In *Proceedings of LREC 2006: Fifth International Conference on Language Resources and Evaluation*, 2006.

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

[62] M. Matessa. Measures of adaptive communication. In *Proc. of the Second Workshop on Empirical Evaluation of Adaptive Systems, 9th International Conference on User Modeling*, 2003.

[63] S. Narayanan, S. Ananthakrishnan, R. Belvin, E. Ettaile, S. Ganjavi, P. Georgiou, and et al. Transonics: A speech to speech system for English-Persian interactions. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2003.

[64] S. Narayanan, P. G. Georgiou, and et al. Transonics: A speech to speech system for english-persian interactions. In *Proc. of ASRU workshop*, 2003.

[65] S. Oviatt. Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, 9:19–35, 1995.

[66] S. Oviatt. User-centered modeling and evaluation of multimodal interfaces. 2003.

[67] S. Oviatt. Human-centered design meets cognitive load theory: designing interfaces that help people think. In *Proceedings of the 14th annual ACM international conference on Multimedia*, 2006.

[68] S. Oviatt, P. Cohen, L. Wu, J. Vergo, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson, and D. Ferro. Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions for 2000 and beyond, 2000.

[69] S. Oviatt, R Coulston, and R. Lunsford. When do we interact multimodally? cognitive load and multimodal communication patterns. 2004.

[70] S. Oviatt, R. Coulston, and R. Lunsford. When do we interact multimodally? cognitive load and multimodal communication patterns. In *Proceedings of the Sixth International Conference on Multimodal Interfaces (ICMI)*, 2004.

[71] S. Oviatt, R. Coulston, and R. Lunsford. When do we interact multimodally? cognitive load and multimodal communication patterns. In *Proceedings of the Sixth International Conference on Multimodal Interfaces (ICMI)*, 2004.

[72] S. Oviatt, R. Lunsford, and R. Coulston. Individual differences in multimodal integration patterns: What are they and why do they exist? 2005.

[73] S. Oviatt, M. MacEachern, and G. Levow. Predicting hyperarticulate speech during human-computer error resolution, 1998.

[74] B. Pakucs. Sesame: A framework for personalized and adaptive speech interfaces. In *Proc. of the EACL-03 Workshop on Dialogue Systems: interaction, adaptation and styles of management*, 2003.

[75] B. Pakucs. Employing context of use in dialogue processing. In *Proc. of CATALOG 04*, 2004.

[76] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *IBM Research Report RC22176 (W0109-022)*, 2001.

[77] O. Pietquin, 2004. A Framework for unsupervised Learning of Dialogue Strategies.

[78] J. E. Pitkow, H. Schuetze, T. A. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. M. Breuel. Personalized search. volume 45, pages 50–55, 2002.

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

[80] K. Precoda and R. J. Podesva. What will people say? speech system design and language/cultural differences. In *Proc. of ASRU workshop*, 2003.

[81] H. Prendinger, J. Mori, and M. Ishizuka. Recognizing, modeling, and responding to users' affective states. In *User Modeling 2005*, 2005.

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

[83] P. Qvarfordt, D. Beymer, and S. Zhai. Realtourist - a study of augmenting human-human and human-computer dialogue with eye-gaze overlay pernilla qvarfordt. In *INTERACT 2005, LNCS 3585*, 2005.

[84] E. Rich. User modeling via stereotypes. *International Journal of Cognitive Science*, 3:329–354, 1979.

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

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

[87] D. L. Roter and J. A. Hall. Studies of doctor-patient interaction. *Annual Review of Public Health 10:163-180*, 1989.

[88] J. Schatzmann, K. Georgila, and S. Young. Quantitative evaluation of user simulation tech-niques for spoken dialogue systems. In Proc. ICSLP, 2002.

[89] K. Scheffler and S. Young. Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning. In *Human Language Technology(HLT)*, 2002.

[90] K. Scheffler and S. J. Young. Corpus-based dialogue simulation for automatic strategy learning and evaluation. In *In Proc. NAACL Workshop on Adaptation in Dialogue Systems.*, 2001.

[91] J. Sheinvald, B. Dom, and W. Niblack. A modeling approach to feature selection. In *The 10th International Conference on Pattern Recognition*, 1990.

[92] J. Shin, G. Georgiou, and S. Narayanan. User modeling in a speech translation driven mediated interaction setting. In *Proceedings of the 1st ACM international workshop on Human-centered multimedia*, 2006.

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

[94] E. Shriberg. Spontaneous speech: How people really talk, and why engineers should care. In *Interspeech 2005*, 2005.

[95] E. Shriberg, R. Bates, and A. Stolcke. A prosody-only decision-tree model for disfluency detection. In *Proc. Eurospeech '97*, pages 2383–2386, Rhodes, Greece, 1997.

[96] N. Y. SOMERS. Ibm delivers new, world class in-car speech recognition system for navigation in 2005 honda cars. http://domino.watson.ibm.com/comm/pr.nsf/pages/news.20040901_speech.html, 2004.

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

[98] W. Wahlster. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, 1 edition, 2000.

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

[100] M. A. Walker, R. J. Passonneau, and J. E. Boland. Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *Meeting of the Association for Computational Linguistics*, pages 515–522, 2001.

[101] W. Ward and B. Pellom. The cu communicator system, 1999.

[102] H. Yan and T. Selker. Context-aware office assistant. In *Proceedings of International Conference on Intelligent User Interfaces*, 2000.

[103] S. Young. Talking to machines (statistically speaking). In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2002.

[104] K. Zechner and A. Waibel. Minimizing word error rate in textual summaries of spoken language. In *Proceedings of NAACL-ANLP-2000*, 2000.

[105] B. Zhou and Y. Je. A hand-held speech-to-speech translation system. In *Proc. of ASRU workshop*, 2003.

[106] V. Zue. Jupiter: A telephone-based conversational interface for weather information, 2000.

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

[108] I. Zukerman and D. W. Albrecht. Predictive statistical models for user modeling. *User Modeling and User-Adapted Interaction*, 11(1-2):5–18, 2001.

[109] I. Zukerman and D. Litman. Natural language processing and user modeling: Synergies and limitations. *User Modeling and User-Adapted Interaction*, 11(1-2):129–158, 2001.