Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 554–571

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Enabling effective design of multimodal interfaces for speech-to-speech translation system: An empirical study of longitudinal user behaviors over time and user strategies for coping with errors☆

JongHo Shin, Panayiotis G. Georgiou*, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory (SAIL), USC Viterbi School of Engineering, Los Angeles, CA 90089, United States

Received 22 June 2010; received in revised form 30 January 2012; accepted 2 February 2012
Available online 11 February 2012

Abstract

The study provides an empirical analysis of long-term user behavioral changes and varying user strategies during cross-lingual interaction using the multimodal speech-to-speech (S2S) translation system of USC/SAIL. The goal is to inform user adaptive designs of such systems. A 4-week medical-scenario-based study provides the basis for our analysis. The data analyzed includes user interviews, post-session surveys, and the extensive system logs that were post-processed and annotated. The annotations measured the meaning transfer rates using human evaluations and a scale defined here called the concept matching score.

First, qualitative data analysis investigates user strategies in dealing with errors, such as repeat, rephrase, change topic, start over, and the participants' self-reported longitudinal adaptation to errors. Post-session surveys explore participant experience with the system and point to a trend of user-perceived increased performance over time.

The log data analysis provides further insightful results. Users chose to allow some degradation (84% of original concepts) of their intended meaning to proceed through the system, even after they observed potential errors in the visual output from the speech recognizer. The rejected utterances, on average, had only 25% of the original concepts. This user-filtered outcome, after the complete channel transfer through the S2S system, is that 91% of the successful turns result in transfer of at least half the intended concepts while 90% of the user rejected turns would have conveyed less than half the intended meaning.

The multimodal interface results in 24% relative improvement in the confirmation mode and in 31% relative improvement in the choice mode compared to the speech-only modality. Analysis also showed that users of the multimodal interface temporally change their strategies by accepting more system-produced choices. This user behavior can expedite communication seeking an operating balance between user strategies and system performance factors. Lastly, user utterance length is analyzed. Longer utterances in general imply more information delivered per utterance but potentially at the cost of increased processing degradation. The analysis demonstrates that users reduce their utterance length after unsuccessful turns and increase it after successful turns and that there is a learning effect that increases this behavior over the duration of the study.

Keywords: Speech-to-speech; S2S; Speech translation; Longitudinal studies; User interfaces; HCI; User behaviors

☆ This paper has been recommended for acceptance by 'Roger K Moore'.

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

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2012.02.001

1. Introduction

Speech is one of the most natural and promising communication modalities for driving *human–human machine interfaces*. The globalization and internationalization of today's world are creating interpersonal interaction scenarios across many domains, such as healthcare, business, and tourism, that are increasingly cross-lingual. Since potential language barriers cause significant communication and access restrictions, the demand for technologies that can help bridge the language gap has grown significantly. Technologies for translation are being developed rapidly but most efforts have been in the field of text-based (machine) translation, such as "Google translate" (http://www.google.com/translate). In the field of speech translation, while the first commercial applications are only beginning to appear now, vibrant research efforts have been underway. These include those at BBN (Kao et al., 2008), CMU (Bach et al., 2007), IBM (Gao et al., 2006), SRI (Precoda et al., 2007), and USC (Narayanan et al., 2003; Ettelaie et al., 2006).

To implement a well-performing and useful speech-to-speech (S2S) translation system, intensive research is demanded along multiple dimensions: from speech recognition and machine translation to interface design (Young, 2002; Knight and Marcu, 2005; Oviatt, 2006). In particular, studies about modeling a user of a speech interface are critical for ensuring wide applications of such a system. Such findings could lead to a speech translation system that can adapt to users in various situations in real time. User-centered systems, in general, that utilize user demographics, cultural information, and user preferences lead to improved usability and satisfaction (Rich, 1999; Krulwich, 1997; Bernstein and Reinecke, 2007; Jannach and Kreutler, 2005).

Most of the user modeling studies in the speech technology community have taken place in the context of spoken dialog systems. Notable user modeling work includes the design and evaluation of multimodal interfaces (Oviatt et al., 2004; Dybkjær et al., 2004; Deng et al., 2004), analysis of user behaviors (Oviatt et al., 2004; Shin et al., 2002), probabilistic user models (Eckert et al., 1997; Zukerman and Albrech, 2001), utility-based models (Horvitz and Paek, 2001), knowledge-based models (Komatani et al., 2003), and user simulation (Levin et al., 2000; Eckert et al., 1997; Scheffler and Young, 2002). It should be noted that mediated interpersonal communication systems (e.g., S2S translation systems) have been used in a very limited way in this context. Early user research with S2S translation systems has been conducted under the Verbmobil project (Bub and Schwinn, 1996) and in our previous work (Shin et al., 2006). Recent advances in S2S systems, however, allow us to conduct further detailed user modeling studies, such as that considered in this work. One goal for the present study is to explore the potential learning effects of longitudinal usage of the S2S translation system. We want to investigate whether the users acquire over time effective strategies to deal with potential sources of uncertainty that eventually can boost the performance. Another goal is to investigate the use of multiple input modalities (e.g., speech, mouse, and keyboard) together and the benefits in facilitating machine-mediated S2S communication. Previous work in spoken dialog systems showed that cognitive load is reduced while using a multimodal interface in comparison with that of a speech-only interface (Oviatt, 2006; Oviatt et al., 2004). In addition, it was reported that multimodal interfaces significantly improved user experience (Deng et al., 2004). Furthermore speech-centric multimodal interfaces provide opportunities for enhanced usability and naturalness and are an increasingly important research direction (Dybkjær et al., 2004; Flanagan, 2004).

In the present study, we set up and performed a scenario-based experiment, in which native speakers of English and Farsi (Persian) interacted using a multimodal interface of an S2S translation system. Three different types of data were collected from the experiment: interviews with participants, surveys, and the log data of the system.

We analyzed the data, both qualitatively and quantitatively, in the following aspects:

1. user satisfaction with the multimodal interface, the S2S translation system, and the experimental setup;
2. level of perceived user proficiency over time in using the multimodal interface of the system;
3. user actions upon successful/unsuccessful interaction turn, with a focus on retry/accept behavior and utterance length;
4. success of interaction, in terms of the number of concepts transferred through the system.

An emphasis of the present study is the consideration of "meaning" as a part of the metric to assess the performance of both the S2S translation system and the related user behaviors. Much like a human translator, the S2S translation system attempts to transfer "meaning" from one language to another language, such as from English to Farsi (Persian) (Narayanan et al., 2003). The process is inherently lossy. Vocabulary words and phrases need to be changed to their

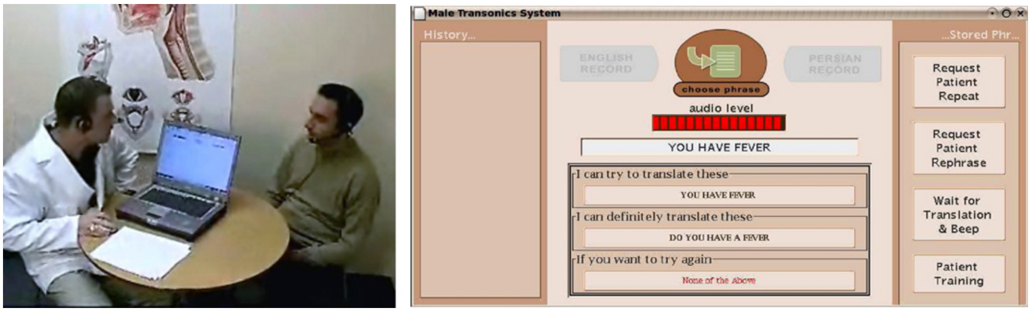


Fig. 1. A scene of conversation between an English-speaking doctor and a Farsi-speaking patient who use the Transonics system for translation. The GUI of Transonics presents a few recommendations to users after the users speak a sentence.

legitimate representation in the target language. However, they will often be re-mapped to more distant equivalents, and grammar and syntax of the target language also typically degrade. As a result, the original meaning will be altered at several different levels (Larson, 1997). It is conveyed sometimes quite closely but more often poorly. Therefore, measuring how well "meaning" is transferred by the S2S translation system becomes important. Existing text translation metrics, such as BLEU (Papineni et al., 2001) and NIST (Doddington, 2002) scores, are based on the comparisons of several human translations with system-produced translations using lexical matching. One recently developed NIST metric utilizes the odds of successfully transferring low-level concepts (Sanders et al., 2008). Likewise, we devised a measure called "concept matching score" in order to measure how well "meaning" is transferred during online conversations of the experiment. Offline log data were utilized for measuring the score. This score refers to the number of concepts in a user utterance (source), which are carried over to the machine-produced utterance (target). Our metric is not intended as a general automated metric but is a human opinion score of the quality of concept transfer.

The paper is organized as follows: the S2S translation system used in our experiments is introduced in Section 2. The experiment set-up and the collected qualitative and quantitative data with descriptions are described in Section 3. The analysis results are presented in Section 4. The discussion is provided in Section 5 and conclusions are in Section 6.

2. System

In this study, we used the Transonics (Narayanan et al., 2003) system designed for mediating and logging spoken conversations between native English and Farsi speakers. Transonics is a two-way translation system with a multimodal interface employing speech, a graphical user interface (GUI), and "push-to-talk" modalities. The application 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. While the system can operate in any domain given the appropriate acoustic and language models; the ones incorporated in the present system focused on the medical domain. To help understand how the system works, Fig. 1 shows a snapshot image from an interaction and a screen capture of the device in a doctor–patient interaction. On the right the processing of the doctor's utterance "you have a fever?" is shown as the ASR output on the top line. The system presents the user two machine-translation options: "you have fever" and "Do you have a fever."

Transonics incorporates a *push-to-talk* interface with which users initiate a speaking turn. This interface has both advantages and limitations; users verify concepts before executing the final decision for 'speaking-out,' but conversations are conducted under a somewhat less spontaneous setting. By design, the interface control of Transonics is also asymmetric in the sense that English speakers (doctors) have exclusive control of the interface, while the Farsi speakers (patients) cannot access the GUI. Two assumptions were made for this design: (1) monolingual patients are assumed to be little trained in using the system and of limited literacy, and controlling the system would not beneficially affect the interaction; (2) even in a monolingual setting, doctors typically lead the interaction; therefore, this asymmetric setting of the interface design is expected to ensure uniform and realistic results from the experiments (Meeuwesen et al., 1991; Onga et al., 1995).

The system decides if the recognized utterance is close enough to a particular utterance or a concept class based on the confidence scores of automatic speech recognition (ASR) and concept classification (i.e., utterance canonicalization). If

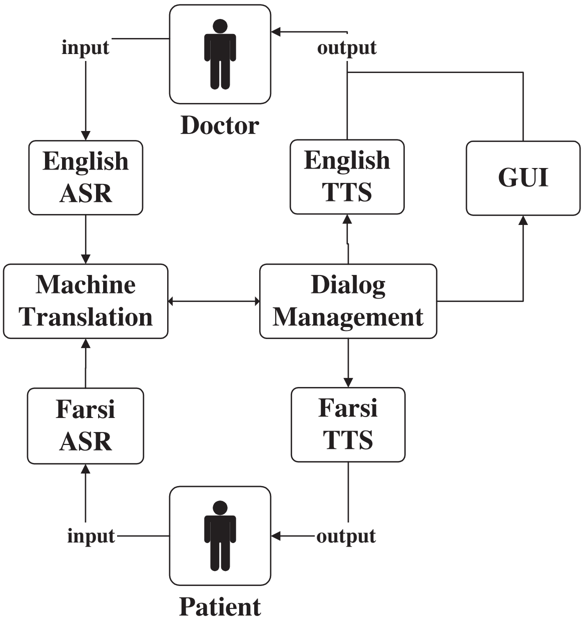


Fig. 2. Simplified data flow diagram of our two-way speech translation system for doctor–patient interactions. English and Farsi automatic speech recognition (ASR) models get the input from users (doctor and patient, respectively) while the machine translation (MT) module is responsible for automatic 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 text-to-speech (TTS) synthesizer.

deemed confident, the cluster-normalized concept form will be transferred to the doctor and, if not, a direct potentially noisy statistical phrase-based translation of the text will be provided. In the choice mode of operation, this decision is left to the user who, as shown in Fig. 1, can select which option to transfer. Most of the time, any incorrect transfer can be detected by the doctor due to the lack of coherence with the discourse of the interaction. Likewise, Farsi-speaking patients can also repeat or repair, verbally or through gestures, when they have detected any incorrect transfer, without controlling the GUI of the system. Note that an experienced doctor, in the case of receiving information that does not match the discourse, can proceed with error control by rejecting the solution provided by the system and requesting additional information.

2.1. Internal components of Transonics

The internal process of the Transonics system involves several components. Fig. 2 shows a simplified block diagram of Transonics with its components. The user's spoken utterance is converted into text 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 study), a process which is inherently lossy, i.e., often the transcript may not accurately represent what the user said. This loss can be due to deletion, insertion, or substitution of spoken words. The output of ASR is further processed by two parallel mechanisms: one by a phrase-based statistical machine translation (MT) module that translates the text from one language to another, and the other by a statistical classifier which attempts to categorize the utterance into one of several predetermined (canonical) concept categories. At this stage the text is mapped from the source language (e.g., English) to the target language (e.g., Farsi). 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 information from one interlocutor to the other.

Fig. 3 graphically shows the details of the system's functional operation and defines the symbols for subsequent clarity. As described above, the MT step operates in 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

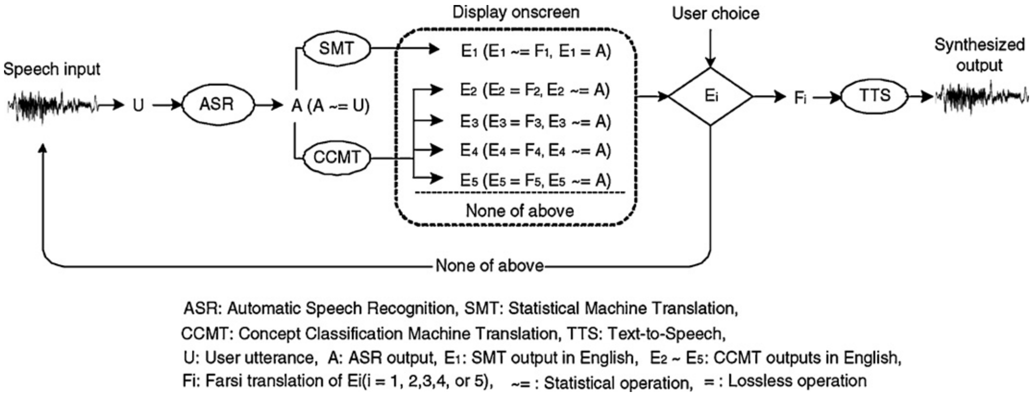

Display onscreen
User choice
Speech input
U
ASR
A (A ~= U)
SMT
CCMT
E1 (E1 ~= F1, E1 = A)
E2 (E2 = F2, E2 ~= A)
E3 (E3 = F3, E3 ~= A)
E4 (E4 = F4, E4 ~= A)
E5 (E5 = F5, E5 ~= A)
None of above
Ei
Fi
TTS
Synthesized output
None of above
ASR: Automatic Speech Recognition, SMT: Statistical Machine Translation,
CCMT: Concept Classification Machine Translation, TTS: Text-to-Speech,
U: User utterance, A: ASR output, 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

Fig. 3. The internal procedure of generating speech translation candidates implemented in the Transonics system. A doctor uses a two-modality interface (push-to-talk) and sees up to five candidates onscreen: one machine translation (MT) candidate (E_1), and up to 4 classifier candidates (E_2–E_5).

has the advantage that it provides a highly accurate back translation since the concepts known by the CCMT were previously manually translated. Thus options E_2–E_5 will be transferred very accurately in the target language, while option E_1 will undergo some further lossy channels.

Detailed descriptions of the symbols in Fig. 3 are as 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 ($U \simeq A \simeq F_i = E_i, \forall\, i = \{2, 3, 4, 5\}$).

2.2. Multimodal interface of Transonics

Multimodal interfaces are considered to be flexible and accommodating to large user differences (Oviatt et al., 2004) and to support richer interactions for the familiar users. In this regard, the interface of Transonics was designed multimodally to fulfill the requirement of quality translation accommodating diverse users.

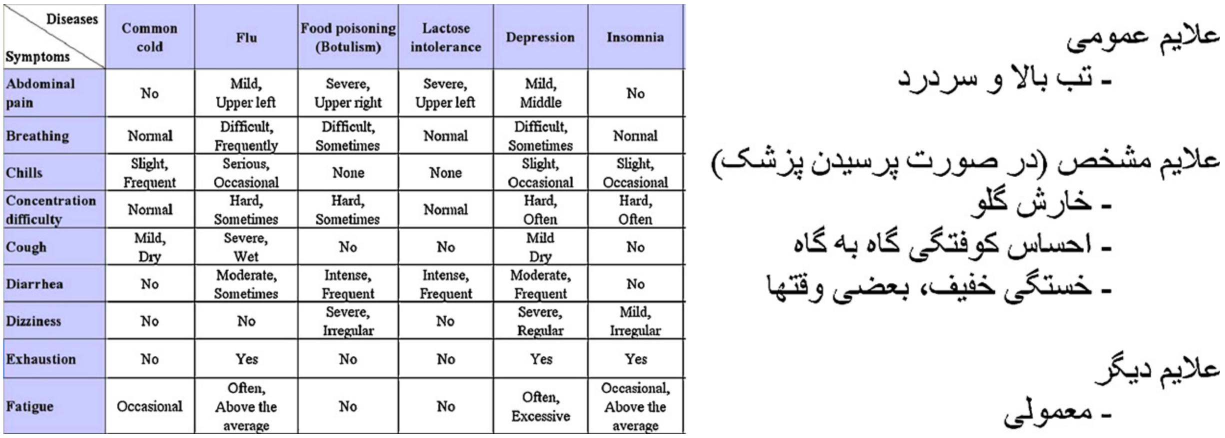
The *push-to-talk* interface of Transonics consists of two input modalities, speech and mouse, to work in conjunction with speech output and GUI, as shown in Fig. 1. After voice input, users are able to make choices given a list of available options through the GUI and are best able to select an appropriate option using the mouse (or touch). Fig. 3 symbolizes this description. U is the input user speech and E_i where $i \in \{1, 2, 3, 4, 5\}$ and "None of Above" are the items of a list available to users to select.

3. Data collection

3.1. Experiment setup

The study is to present analysis results of cross-lingual conversations, mediated by the Transonics system, between English-speakers playing the role of doctors and Farsi-speakers playing the role of patients. Furthermore, since this data collection was performed over several interaction sessions with the same users, it allows us to investigate the learning effect of users and varying user behaviors across sessions during experiments. Note that human adaptation to the mediation device is regarded in a sense that the mediation system directly interacts with humans using the form of visual and audio representation.

For the purpose of study, we designed experiments conducted over 4 weeks with four English–Farsi interlocutor pairs. The age range of the participants was from 20 to 30 years, and they were graduate and undergraduate students at USC. A 1-h training session was given to all participants before the actual experiments. Details of the training session included how to do the experiments with the given scenarios when using Transonics. In this 1-h training session, 30 min were devoted to interactive instruction. The next half hour was a verbal explanation of the experiments. No special explicit training for the usage of the speech interface was, however, given to the participants in this training session.



Symptoms \ Diseases	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

Fig. 4. Simplified example reference documents of the role-play experiment for doctor-role English speakers and patient-role Farsi speakers. On the left, a sample diagnosis reference for doctor-role participants is illustrated, in which six diseases (in the columns) are presented with several symptoms (in the rows). On the right, a patient card indicating *common cold* with a few symptoms is presented, which is written in Farsi for patient-role participants.

In actual clinical encounters, it is difficult to collect conversation data between English-speaking doctors and Farsi-speaking patients using an experimental technology system. Hence, to collect representative doctor–patient interactions with minimal domain error but sufficiently realistic domain language usage, a role-play experiment was designed with native speakers of American English playing the doctors' roles and Farsi speakers as patients[1]. Such a role-play approach is widely used in medical education and training notably with the so-called standardized patient approach and in speech technology evaluation (Walker et al., 2002; Belvin et al., 2004; Barrows, 1987). Each role was prepared with reference documents to be used during the interaction. In particular, the doctors had three documents for each experiment session, namely a diagnosis reference, a disease treatment reference, and a medical term dictionary. The diagnosis reference is a table of 12 diseases (common cold, flu, food poisoning, lactose intolerance, depression, insomnia, hypertension, high cholesterol, liver cancer, lung cancer, SARS, and diabetes) by 30 symptoms. The 30 symptoms vary depending on the disease. We developed the diagnosis references in a realistic fashion by using information from MedicineNet (2006). The other two documents are a disease treatment reference and a medical term dictionary: The former includes detailed treatments for each disease, with which doctors can provide treatments to patients at the end of diagnosis sessions. The latter is a document of disease/symptom definitions that are helpful for both doctor and patient participants in understanding difficult medical terms. The patients, on the other hand, have access to only two documents: a medical term dictionary and a symptom card with four symptoms written in Farsi. Different sessions employ different diagnosis references (for doctors) and symptom cards (for patients). An example diagnosis reference for a doctor and a symptom card for a patient are shown in Fig. 4.

3.2. Experiment procedure and data collection

Each English-speaking doctor and Farsi-speaking patient pair conducted eight different conversation sessions during 4 weeks, performing two sessions per week. Two sessions were carried out on the same day with a 20-min break time between sessions. In total, there were 30 min per session, 10 min for oral interview, for a total of 1 h and 40 min per day. Throughout the whole study each pair handled eight different scenarios across these eight interactions. In each session, the doctors attempted to identify the patients' diseases. Neither party had any information on the nature of the information the other had. All four pairs of subjects followed the same sequence of scenarios: common cold, liver cancer, food poisoning, SARS, hypertension, lung cancer, insomnia, and depression. Another assumption was that the difficulty level of the scenario is constant due to the open-ended questions and the common medical knowledge used in the design of topics. This assumption helps reduce deviation from the normalized results. All participants

[1] For simplicity the doctor-role English speakers will be referred to as doctors and the patient-role Farsi speakers as patients in the rest of the paper.

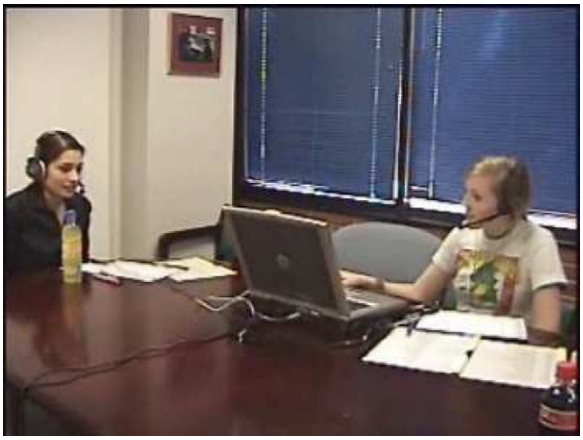


Fig. 5. A snapshot of a diagnosis session using Transonics. A doctor-role native English speaker (right) controls the device for the session and performs a diagnosis with a patient-role Farsi speaker (left), based on the given documents for a specific scenario.

provided comments after each session regarding the difficulty level of the topic, but there was no topic that the participants identified as "most difficult" among the eight scenarios. The experimental scenarios allow for significant freedom of interaction but provide a guideline that minimizes unknown factors, such as participant domain knowledge discrepancy, while the experimental setting was kept consistent throughout the study. All participants used headsets and were instructed only to interact through the device to ensure information was only being transferred through the device. In this particular setting of audio masking effect, the participants only hear translated audio voice through the device. For minimal interruption, the experimenter left the room during the experiment and notified the participants when the interaction time limit was reached.

Fig. 5 shows a snapshot of a diagnosis session where an English speaker (right) is interacting with a Farsi speaker (left). Each session is mostly a question–answer conversation based on the given documents for one of eight scenarios. Although the English speaker actively controls the Transonics system (due to the nature of the chosen scenario), the other speaker (of Farsi language) was observed to sometimes take initiative in conversation by gesture or speech. For the data analysis, not only were logs of the experiment sessions collected, but oral interviews were performed with each pair for 10 min before and after each session to collect pre- and post-session surveys. Demographic and general questions were answered by participants before the collections started. Each session was video-recorded so we could later observe and analyze the interactions that unfolded during the experiment.

3.3. Data collected for analysis

For the study, we analyzed conversation and user opinion data collected from different sources, including user interviews, surveys, and the log data. This data collection was designed in such a way as to minimize variability of experimental conditions by keeping the system, experimental location, and experimental settings constant. Details of the procedure and the type of data collection are as follows:

1. A 10-min user interview with the two participants was conducted by the experimenter right after each session. The same person administered all 32 interview sessions. Questions were general and open-ended, such as "How do you feel about the session you just finished?", and interviewers did not guide the participants in any response, allowing the interviewees to freely express their opinion. The most often discussed topics were system performance, personal concerns during the experiment, and suggestions for an updated system design;
2. There are three types of surveys collected: first, a one-time survey given to participants before the whole experiment; second, a survey given before each session so each participant completed out this four times in total; third, a survey given after each session (completed four times in total by each participant). The first one-time demographic survey included assessment of users' general technology proficiency, demographic information, and experience with multimodal interfaces. The second pre-session survey included questions regarding that particular day's feelings, speech interface experiences, and any changes in personal experience compared to the previous session. This

Table 1
Table shows a simplified portion of the data log acquired automatically by running the Transonics system. There are system routing tags (FADT, FDMT, FMDT, FDGT, FDGC, and FGDT—F: flow; A: audio server; D: dialog management; M: machine translation; G: graphical user interface; T: text; 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, includes time stamps, utterance sequence, confidence, and class numbers.

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

pre-session survey was designed for the measurement of any changes with regard to the user and his/her experience of the speech interface. The last post-session survey included questions about user satisfaction, perception of the overall system performance, difficulty level of topic or using the system, and any open suggestions. This post-session survey was designed to target gathering of user opinions about any improvements in or decrements of user satisfaction while participating in the experiment and using the system.

3. The Transonics system is also equipped with a logging mechanism, so all the spoken conversations of the experiment are logged in the text and audio format. In this regard, all 32 sessions' logs were collected. These logs contain not only principal user actions and system status information during the session but confidence levels of the machine-produced utterances, user selected items, recognized hypotheses of the ASR component, translated hypotheses of the translation components (from both the statistical machine translator, SMT, and the concept classifier machine translator, CCMT), audio-recorded user voices, and synthesized system voices in text. Table 1 shows sample 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 come with the system routing tag.

3.4. *Transcription and annotation of logs*

Among the three types of data collected from the experiment, the log data contain diverse information in text form and can be automatically processed. The log data were analyzed twice: first, we examined 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 scores that indicate how many concepts are transferred successfully from the original user utterance to the targeted machine-produced utterance. The annotated logs became the main source for the subsequent analysis in this study.

3.4.1. *Concept matching score*

As introduced in Larson (1997), meaning can be the most important metric in evaluating speech translation results. In reality, we do not expect perfect word-to-word literal translations. Instead, we check the translation results to see if enough meaning is transferred from the original to the target. Inspired by this idea, we devised a metric called "concept matching score (CMS)" to assess the transferred meaning through the Transonics system. The idea of the CMS was borrowed from the Linguistic Data Consortium's human assessment metrics (Ma and Cieri, 2006). In particular, Ma and Cieri (2006) mention: "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." In the study, we assign CMS to utterances based on the number of concepts in the utterances—how many concepts are transferred from the original utterances (source) to the target utterances (destination), either through translation or within the same language, e.g., through a lossy speech recognition channel.

The CMS scores were manually assigned by humans 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 depicted in Fig. 3.

Measuring correct meaning in utterances requires linguistic skills which can be found only in native and culturally aware speakers. To assign CMS scores to the utterances of the log data, we hired 4 bilingual annotators fluent in both English and Farsi. The total number of investigated utterances – from English and Farsi speakers – was 2435, and two annotators took charge of transcription of utterances and the other two in assigning scores to utterances. To make the data manipulation objective and consistent, 2 h of training and calibration sessions were given to each of the 4 bilingual speakers, in which they were provided with verbal instructions and a few samples of transcription and CMS assignment. The CMS score is assigned pair-wise to utterances based on the following guidelines:

Guideline for the concept matching score assignment	
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 understand the whole meaning.
0.2	Few concepts are transferred, such that users will rarely understand the whole meaning.
0.0	None of the concepts are transferred.

3.4.2. *Reliability of the concept matching score assignment*

One way of verifying how well the CMS was assigned to an utterance or identifying any critical bias in the assigned CMS is to measure the percentage of agreement between the two CMS annotators. To measure this agreement between the two, we acquired two sets of CMS assigned to all the pairs of the same-path utterances. The total number of the common processed entries by each was 6353. For the purpose of this testing, the agreement percentage was measured by Kappa Coefficient, and the value of Kappa was 0.52. This Kappa value indicates moderate strength of agreement according to the common interpretation of Kappa Statistics.

4. Results and discussion

In this section we present results based on the analysis of the interviews, surveys, and annotated logs. We incorporated a qualitative and quantitative data analysis investigating various dimensions. The qualitative methodologies include analysis of the videos, audio logs, and in-person user interviews. The quantitative methodologies include a number of statistical tests on the log data and analysis of the post-session surveys. Particularly the statistical tests include analysis of user behaviors and system actions. The log data used for the statistical tests contains user actions, user utterances, and machine-produced utterances that were logged during conversations through the Transonics system, as well as the annotations (e.g., CMS and human translations). We employed the statistical analysis toolkit SPSS 15.0.

4.1. *Qualitative data analysis*

4.1.1. *User interview*

The participants in the experiments, after each session, had an opportunity to verbally provide not only comments on the conducted conversation but also on the performance or usability of the system and the speech interface. This user interview session took about 10 min.

According to the comments in the interview sessions, the participants had managed to conduct conversations successfully in most of the cases. However, their minor opinions implied they had spent a substantial amount of time and effort in handling errors. These errors include speech recognition errors, translation errors, and dialog errors. And these errors eventually caused incorrect concept deliveries from one participant to another. To overcome any errors in the conversation the participants utilized a few strategies, such as repeat, rephrase, change topic, and start over, and believed that they had effectively handled the error situations using these strategies. Some quantitative analysis results

Table 2
Summary of overall statistics from the survey data of doctor-role English speakers that were collected after each interaction session finished: user satisfaction, user perceived difficulty when using the interface, user adaptation to using the system, and user-perceived system performance. The two sets of numbers are mean and standard deviation.

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 3
Perceived user performance during the 4 weeks of the experiment, in terms of using and learning the functionality of the Transonics system. The range of performance level given to the participants is (1: "very bad"; 7: "very good"). The two sets of numbers are mean and standard deviation.

	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)

about the strategies in error conditions are presented in our previous work (Shin et al., 2002) in which user log data of a spoken dialog system were analyzed and a conditional probability model smoothed by weighted ASR error rate was proposed.

One interesting observation from the interview sessions is user adaptation to errors over time. The participants reported that they became more comfortable dealing with any type of errors as they experienced more conversations using the system. Especially, most of the participants commented that their adapted strategies in the last session (week 4) were effective in dealing with errors. The user-adapted strategies are analyzed using the log data and presented in the later section.

4.2. Quantitative data analysis

The surveys and system logs – and their subsequent annotation – provide rich data for detailed quantitative analysis.

4.2.1. Surveys

Analysis of surveys gives us more insightful results about the user and his/her experience during the experiment. In the study, we collected data using three different types of surveys as introduced in Section 3.3: first, the collection of general knowledge about participants; second, the status of participants before each experiment session; third, user experiences and opinions about the conducted session and the system they used. In the following results section, only relevant analysis results on the selected questions are summarized.

Technology background: The analysis of the demographic surveys provides variances of user background in the technology aspect. According to the collected data, none of the participants had any experience with speech-enabled systems before the experiment. The term, "speech-enabled system" represents any computing devices/applications equipped with speech recognition, such as a speech translation application or a call center spoken dialog system. Some details about users and backgrounds are the following: the average level of proficiency in general technology (1: comfortable; 7: never comfortable) was 3.0 (std. 1.4) from four English-speaking participants and 2.5 (std. 1.0) from four Farsi-speaking participants. The average level of proficiency in dealing with computers (1.0: better than most; 5: worse than most) was 2.25 (std. 0.95) from the English-speakers and 1.25 (std. 0.5) from the Farsi-speakers.

The analysis of the pre/post-session surveys was performed on the data collected from English-speaking participants. The asymmetric design of the interface of Transonics prompted focus more on the doctors who control the system during the experiments. On the other hand, patients only used the speech interface, which limits information in many aspects, such as effects of the visual modality and dialog options. The analysis of the pre-session surveys – that attempted to gauge the status or state of the participants before each session – did not provide any significant evidence in differences among the participants. The following results in Tables 2 and 3 are from the analysis on the post-session surveys administered at the end of each session.

Satisfaction: We attempted to measure satisfaction of the users with the overall system and interface. This helps us to understand and address deficiencies in the system design and implementation. Table 2 summarizes the analysis from the post-session surveys; we investigated user perception on overall satisfaction, difficulty in using the system, adaptation to the system, and overall system performance.

Self-performance: Another interesting question regarded how users perceived their performance over the 4 weeks of the experiment. We define user performance as the level of proficiency in employing the system. For qualitatively measuring performance we employed the post-session surveys over the 4 weeks. To minimize the effect of system performance on users' self evaluation of their performance, users were instructed to disregard the system errors in their self-evaluation. Table 3 indicates how the participants perceived their performance over the 4 weeks of experiment, using the range of (1: "very bad"; 7: "very good"). The figures show that the participants believed they were getting better in their performance overall, despite a performance drop in the 4th week compared to week 3. With regard to user performance over time, additional analysis results using quantitative data are presented in the following sections.

4.2.2. *Distribution of meaning transfer rate in successful or unsuccessful turns of the annotated system logs*

The following analysis results are presented based on the data of the doctor-side interactions as more meaningful analysis results are expected. The annotated system logs and video data are primary resources for the analysis. The analysis results of the patient-side interactions would be noted separately.

Statistics of the collected data: The total number of the analyzed English-spoken utterances was 1489 which includes both the number of collected utterances during the experiment and the English translations from the patient utterances. Additionally, for the English speaker the average number of utterances per session was 46.5 (std. 16.7); the average user acceptance rate per session was 0.64 (0.09), representing that users accepted 64% of the machine-processed utterances during the experiment using the choice mode system; the average number of words per utterance was 5.16 (2.0); and the average concept matching score between a user utterance and a system choice that was accepted 0.84 (0.19), indicating that users accept some degradation (average 16%) as a preferable option to retrying or rephrasing. On the patient-role Farsi speakers side, the total number of utterances was 946 which includes both the number of Farsi utterances and the translations of the doctor utterances; the average number of utterances per session was 29.6 (std. 11.8); the average number of words per utterance was 3.34 (2.3); and the average concept matching score between user utterance and translation was 0.56 (0.39), indicating that on average 56% of the original concepts were transferred through the system. Note that Farsi speakers' speech gets automatically translated and they do not have the option of aborting or choosing or any other access to other modalities; therefore all their speech input will get translated.

Accepted meaning degradation: With the multimodal interface of Transonics, doctor-role English speakers can accept or reject machine-processed utterances based on their assessment of the speech recognized text and the classifier proposed classes. We wanted to identify how much degradation users were willing to accept and how this varies across doctors. Based on our analysis, there is a person-specific accept/reject threshold that can be employed for recommending feedback in error conditions. We define a turn in which the user accepts one of the system proposed utterances as a successful turn. Using the concept matching score (CMS) from the annotated logs, we can see in Fig. 6 the distributions and cumulative sum graphs of meaning transfer (CMS) from the user utterance to the selected machine-proposed option for both successful and unsuccessful turns. The successful turn is represented as (U, E_i), where $i \in \{1, 2, 3, 4, 5\}$, and the unsuccessful as (U, E_i), where i =*None of Above*, using the symbols introduced in Fig. 3. Note that this represents the degradation due to the ASR alone and the ASR and concept classifier. The average meaning transfer rate in user selected successful turns was 0.84 (std. 0.19), which represents 84% correct meaning transfer rate. Unsuccessful turns on the other hand have a 0.25 (std. 0.21) meaning transfer and make rejection necessary. The same results analyzed per individual doctor were 0.83 (0.24), 0.86 (0.15), 0.80 (0.2), and 0.87 (0.14) for successful turns and 0.33 (0.25), 0.20 (0.18), 0.29 (0.22), and 0.21 (0.16), respectively, for unsuccessful turns.

Boosted meaning transfer rate with the multimodal interface: Utilizing multiple modalities plays an important role in the design of an interface for interaction-oriented systems, especially when we incorporate speech as one of the modalities. For example, it was reported that a multimodal interface enhanced the overall system and user performance by judiciously adopting multiple modalities, as introduced in the studies (Oviatt et al., 2004; Deng et al., 2004). Thus, in addition to speech, we incorporated a GUI and a "push-to-talk" modality into our system design. In this section we investigate the benefits arising from the multimodal nature in terms of improved concept transfer.

Using the annotated data and assigned CMS scores, we compared the CMS of user utterances and the corresponding translations under three conditions: (1) multimodal interface, with user choice provided by the GUI; (2) multimodal confirmation interface, with the user confirming the utterance to be translated; (3) unimodal, speech-only interface.

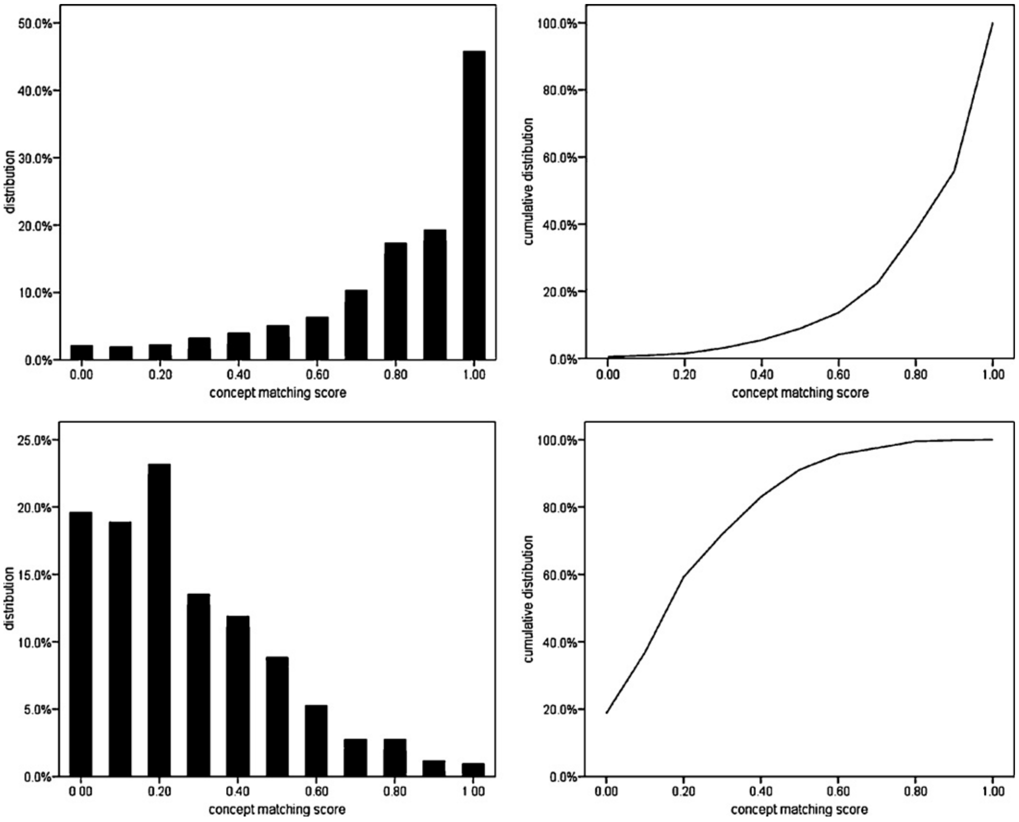


Fig. 6. (Top left) CMS between user utterance and the corresponding machine-processed utterance which is accepted by user (between *U* and E_i where $i \in \{1, 2, 3, 4, 5\}$ as in Fig. 3); (top right) cumulative CMS between user utterance and user accepted machine-processed utterance (successful turn); (Bottom left) CMS between user utterance and user rejected machine-processed utterance (between *U* and E_i where *i* = *None of above* as in Fig. 3); (bottom right) cumulative CMS between user utterance and user rejected machine-processed utterance (unsuccessful turn).

Table 4
Concept matching score and its standard deviation between user utterance and the corresponding translation under unimodal and multimodal interface settings with or without a filtering option.

	Concept matching score (CMS) between user utterance and the corresponding translation		
	No filtering option	Binary options	Multiple options
Unimodal	0.51 (0.33)		
Multimodal		0.63 (0.25)	0.67 (0.23)

A multimodal choice interface is the equivalent of the path CMS[U,F_i] (as in Fig. 3) where $i \in \{1, 2, 3, 4, 5\}$ and *i* is chosen by the user through the GUI and the user has the choice to reject all options. In the multimodal confirmation interface *i* corresponds to the machine-selected best choice and the doctor can only accept or reject. The third case of the unimodal interface is the equivalent of the path CMS[U,F_i] where $i \in \{1, 2, 3, 4, 5\}$ and *i* corresponds to the machine-selected best choice and no rejection can take place.

Table 4 presents the successfully transferred 51% meaning of the original user utterance by the unimodal interface. When other modalities are employed there is a significant 24% relative improvement through the use of the confirmation mode to a 63% CMS and a further improvement through the choice mode to 67% CMS, a 31% relative improvement.

Longitudinal study of varying user/system performance over time: According to the results of our surveys above, the participants perceived an increased self-performance over time during the experiment (Table 3). The improved performance may be investigated from the system logs, e.g., through the speed of communication, the amount of retries, etc.

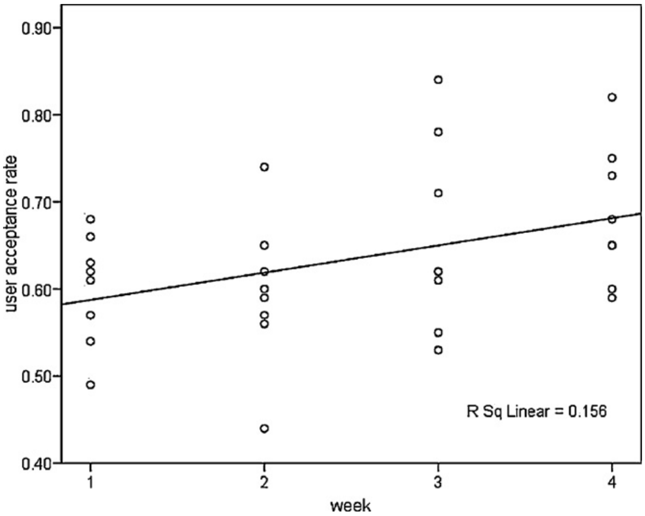


Fig. 7. The linear trend represents an increasing user acceptance rate during the 4 weeks of the experiment. Each participant performed eight conversation sessions over 4 weeks (two per week), and the figure depicts the user acceptance rates of 32 sessions collected by four participants. Each session's user acceptance rate is represented as a circle. In the experiment, the doctor-role English speakers control the multimodal interface of Transonics and accept or retry the machine recommendations for translation to communicate with patient-role Farsi speakers.

Table 5

Meaning transfer rate in ASR output and in user-accepted machine-processed utterance. The meaning transfer rate in ASR output is formulated as CMS (U, A). The meaning transfer rate in the user-accepted machine-processed utterance is formulated as CMS (U, E_i where $i \in \{1, 2, 3, 4, 5\}$). The meaning transfer rate in user-accepted machine-processed utterance differs from that in ASR output in a sense that it varies and results in a significant difference over time according to ANOVA.

	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)

We first observed the longitudinal accept/retry behaviors of users from the log data. Fig. 7 shows an increasing user acceptance rate over time, indicating a trend of increasing user performance. This increasing user performance expedites effective communication in terms of the communication speed. The user acceptance rate in a session is defined as the ratio of the number of user-accepted machine-processed utterances options over the total number of user utterances in the session.

System performance: In the results of Fig. 7, we assumed that the Transonics system is consistent in its performance because of the static state of its models and the same interface being used during the experiment. Also, before the experiment, the participants were trained enough to be comfortable with usage of the Transonics system. Therefore, the initial levels of users' skills in dealing with the system were normalized for the experiment. In reality, however, many system performance factors affect user behaviors, for example, varying speech recognition precision. In this regard, we investigated the system performance and its relation to user accept/retry rates. The factors considered in the following analysis are ASR recognition errors, meaning transfer rate, and translation quality.

ASR performance: The ASR output represents the front-end system that transforms the audio input format into text. Based on this text output, the system and/or user perform the error handling and input to the translation module. Table 5 presents the meaning transfer rate, CMS (U, A), of the ASR over 4 weeks. We can see no significant difference over time (ANOVA: $F = 2.0$, $p = 0.12$) in the CMS of U, the user utterance, and A, the ASR text output.

User's intended transfer rate: We measured meaning transfer rates, CMS (U, E_i) where $i \in \{1, 2, 3, 4, 5\}$, based on user selection i. This metric is in effect what the English user, lacking any understanding of the target language, believes the concept translated will be, and it does not include all the errors in the translation path. Table 5 presents the decreasing meaning transfer rate over 4 weeks. The significance ($F = 4.17$, $p < 0.01$) has been confirmed by ANOVA. Further analysis results using a post hoc test with Tukey HSD confirmed that the meaning transfer rate of the first week

Table 6
Average utterance length and its standard deviation in three conditions: overall, after successful turn, after unsuccessful turn. The utterance length is defined as the number of words in the utterance.

	Overall	After successful turn	After unsuccessful turn
Average utterance length	5.16 (2.0)	5.26 (2.1)	5.0 (1.9)

Table 7
Percentage of reduced utterance length after unsuccessful turns indicated the trend of learning effect in user strategy. The statistics were measured from the log data of English speakers who control the multimodal interface of Transonics.

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

differs dramatically from that of the last week. The result in the second case prompted a more thorough investigation of the reason for this CMS reduction as a reduced meaning transfer rate can be counter to the system's objective. We investigated the actual log data line-by-line to identify the cause of this and concluded that through experience users become accepting of functional words and make faster progress in communication as they use the system more. Note that incorrect placement of function words appear to be errors in the aspect of grammar, less likely of semantics. For instance functional words can damage the CMS score but have little impact in allowing the patients to reconstruct the original concepts, such as in the example utterance "AND DO YOU HAVE ANY OTHER SYMPTOMS," in which the word "AND" is an insertion that lowers the CMS.

Overall transfer rate: The translation result is considered in relation to user behavior. The trend of increasing user acceptance rate (Fig. 7) and the decreasing CMS (Table 5) in the user-accepted utterances could result in decreased translation quality. In this regard, we investigated the translation quality over 4 weeks: 0.69 (0.22), 0.64 (0.22), 0.66 (0,21), and 0.65 (0.24). The ANOVA measure on the translation quality confirms that there is no significant difference (ANOVA: $F = 1.3$, $p = 0.17$). The meaning transfer rate in translation was measured between user utterance and the corresponding translation, formulated as CMS between U and F_i where $i \in \{1, 2, 3, 4, 5\}$, where i is a user choice. The constant translation quality over time and the trend of increased user acceptance rate over time together imply that users are adapting and training themselves to achieve their goals.

4.2.3. User strategy on errors by utilizing utterance length

Utterance length: The participants in the experiment who used the multimodal interface of Transonics demonstrated effective strategies, especially when dealing with errors. Users reduced the length of their utterances and rephrased the previously failed utterance to get back on track from the error situations. The application of this user strategy became more apparent when there was a chain of consecutive errors. In contrast they attempted to speak longer sentences using the multimodal interface when they faced a normal flow of conversation. Table 6 shows the overall statistics about (English) utterance length measured after user accept/retry behavior during the experiment. The utterance length is measured in number of words. Table 6 indicates that the users spoke relatively longer utterances after successful turns and spoke in shorter utterances after unsuccessful turns, in which users retry (5.0 words in average).

Longitudinal utterance length variation: This user strategy contributed to lowering the overall error rate; however, the amount of information exchanged may be lowered too. The figures of Table 6 suggest patterns of user behavior indicating that users seek equilibrium in their behaviors during 32 conversation sessions by modifying the length of their utterances. According to the user interview data and the audio/video data, doctor-role English speakers utilized the utterance length judiciously to achieve the ultimate goal – the correct diagnosis – of each experiment session within the 30 min.

They dramatically reduced the utterance length when they faced a difficulty in dealing with errors; however, once they got a short utterance recognized by the system, users spoke the same length or the longer sentence. From the log data, we observed that users spoke at least at their average length or longer after a successful turn at 62%, 62%, 67%, and 68% of the time over the 4 weeks, respectively.

Another important aspect considered in the study is learning effects. Over time, the users exhibited a behavior of utilizing shorter utterances as an effective strategy to overcome errors, although in general otherwise they employ

Table 8
Average utterance length and its standard deviation collected from the offline log data of 32 interaction sessions conducted over 4 weeks by four English (E) and four Farsi (F) speakers.

	E/F user 1	E/F user 2	E/F user 3	E/F 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)

longer utterances. Table 7 shows the percentage of reduced utterance length after unsuccessful turns during the 4 weeks of the experiment. It provides a trend of temporal learning of user strategy. Note that users used the strategy of reducing the utterance length more during the last 2 weeks compared to the first 2 weeks.

5. Discussion

The purpose of this study was to identify and analyze metrics that can affect S2S system performance and user experience and guide future developments. A well-designed speech-to-speech translation system needs to deal with different user behaviors according to the appropriate models and the interaction goals.

We observed explicit user differences in their overall utterance length from the collected data of the study, in which four English and four Farsi speakers converse using the multimodal interface of Transonics. ANOVA measure confirms that there is a significant difference in the utterance length of participants: $F = 48.7$, $p < 0.01$ (English speakers), and $F = 90.7$, $p < 0.01$ (Farsi speaker). Table 8 shows the difference of utterance length between participants. English/Farsi speaker 1 and English/Farsi 4 had shorter utterances compared to English/Farsi 2 and English/Farsi 3. Surprisingly, the conversation partners seem to have the same tendency of managing utterance length, synced by each other. For example, English- and Farsi-speaking partners with user numbers 2 and 3 have longer utterance lengths overall, compared to English and Farsi partners with user numbers 1 and 4.

One of the hypotheses in the study was that translation quality improves as users become more familiar with the system. However, the study showed that there was no significant improvement or degradation in the translation quality in terms of meaning transfer rate over time. This can be due to the combination of different user behaviors which vary over time and the responses of the system to the varying user behaviors. For example, the quality of translation gets improved when users speak shorter utterances, but it is degraded by increased acceptance of functional words. Also, the system modules, such as acoustic or language models, have been kept static throughout the study; therefore, we expect statistically consistent system performance. This points to a need for dynamically adapting the system to the user, so the system can better cope with changing user behaviors, and improving the underlying machine models. For example, a well-designed speech-to-speech translation system would provide more information when users shorten the length of speech in error conditions (refer to the result of Section 4.2.3).

Another discussion point in the study is the possible bias in the assignment of concept matching score (CMS). Concept matching score is a subjective metric assigned by humans, although we used 4 bilingual annotators to cross-validate the two sets, and we trained and attempted to standardize the process. We prepared the annotators extensively, and we gave many standardization examples and also averaged the CMS scores between the two sets for the analysis for the statistically large pool of data. Future work can include more unbiased concept matching scores.

6. Conclusion

In this study, we designed a scenario-based experiment performed over a period of 4 weeks by English speakers playing the role of doctors and Farsi speakers playing the role of patients. The experiment participants conversed with each other to achieve the goal of the designed conversation task. Not only quantitative log data of the system and surveys were collected but qualitative data, such as audio/video recordings of the interactions and user interviews, were also gathered. The diverse source of collected data provided us opportunities to investigate various aspects of user behaviors and system functions. Analysis results using both quantitative and qualitative data were reported in the study; notably, the log data collected over 4 weeks contributed to the quantitative analysis in the longitudinal study of user behaviors. Qualitative data, such as from user interviews, furnished ideas for the design of this longitudinal study. The analysis results provide details of user/system performance, how users react to errors, and how well users utilize and learn the multimodal interface of Transonics over time. The analysis of user interviews brought up the issue of effective user strategies coping with errors, such as repeat, rephrase, change topic, and start over. Further analysis revealed a learning effect during the 4 weeks of the experiment, in which users got comfortable in dealing with errors as they made more use the multimodal interface of the Transonics system.

The analysis of quantitative data provided more concrete support for the conclusions about varying user behaviors, system performance factors, and the combinations of both. Meaning transfer rate is an important metric for the measurement of the system components' quality, which is useful for improving the design of the overall speech-to-speech translation system. In this regard, a novel measure, the Concept Matching Score, was devised in the study. Concept matching scores were assigned by human evaluators based on the number of concepts in the utterances, i.e., how many concepts are transferred from the original utterances (source) to the target utterances (destination). Four bilingual speakers assigned scores to all the available utterances, and the scores were cross-validated.

The survey data analysis provided basic statistics about user backgrounds; in particular, participants' level of proficiency/knowledge about technology was above average, but none of the participants had experience with speech-enabled interfaces before the experiment. Based on the survey results, we expected that the training sessions for all the participants standardized the participants' handling of the multimodal interface. Another aspect of analysis from the surveys was about participants' experiences during the 4 weeks of the experiment. They reported above average user satisfaction and system performance, easier use of the multimodal interface, and easier adaptation to using the Transonics system as well as an increasing rate in user-perceived performance.

The first finding of this analysis provided statistics regarding how much meaning is transferred in successful and unsuccessful turns when users use a multimodal interface. This result indicates a threshold of user acceptable or unacceptable system performance for successful or unsuccessful turns, in terms of meaning transfer rate. Average meaning transfer rate in successful turns was 84% and in unsuccessful turns was 25%. In addition, 91% of successful turns resulted in more than 50% meaning transfer, and 90% of unsuccessful turns resulted in less than 50%, an observation that suggests that the users are very effective in their filtering of bad options through the ASR output.

The second finding was about the benefit of incorporating a multimodal interface into the design of a speech-to-speech translation system. It was reported that the meaning transfer rate acquired through a multimodal interface setting is higher in comparison with the meaning transfer rate acquired through a unimodal speech-only interface setting. A 24% relative improvement was reported by adding just a visual cue to the speech only interface, and a 31% relative improvement was observed by adding up to 5 visual cues to the speech only interface.

The next finding was about varying user strategies and system performance factors over time. This result plays an important role in the design of a speech-to-speech translation system because some dynamic adaptation can be applied to the conventional static design of the system, utilizing the information of varying user behaviors. It turned out in our study that users showed a trend toward an increasing rate of acceptance over the 4 weeks of the experiment. The underlying components, ASR, translation modules, etc., were kept constant and performed consistently in terms of both their metrics (WER, BLEU) and the meaning transfer rate, but users gradually increased the average number of accept behaviors in relation to rejections. This resulted in a lowering of the meaning transfer rate over time. We conjecture that users got accustomed to the usage of the system so they tried to make the conversations smoother and faster by ignoring some functional words that do not affect the final translation quality. Another study could investigate longer term usage and identify whether the users reach some more appropriate balance of acceptance/rejection as their learning improves.

Lastly, user utterance length was analyzed in two cases: after successful turns and after unsuccessful turns. Results show that users relatively increase their utterance length more after a successful turn, compared to that after an unsuccessful turn. The result of the learning effect with regard to the two cases indicated a trend toward an increased usage of this strategy by the users (reducing utterance length after an unsuccessful turn) as their familiarity with the system increased. The user utterance length plays an important role in the design of a speech-to-speech translation system because longer utterance lengths mean the delivery of more information but also higher probability of statistical processing errors and vice versa in the case of the shorter utterance. A better design of a speech-to-speech translation system would take advantage of utilizing the length of user utterance judiciously, which can result in the optimal delivery of information for cross-language communication.

Acknowledgements

This research was supported in part by the National Science Foundation, award number 0911009, the Babylon/CAST program, contract N66001-02-C-6023, and by the DARPA TransTac program, contract number NBCH1050027.

References

Bach, N., Eck, M., Charoenpornsawat, P., Köhler, T., Stüker, S., Nguyen, T.L., Hsiao, R., Waibel, A., Vogel, S., Schultz, T., et al., 2007. The CMU TransTac 2007 eyes-free and hands-free two-way speech-to-speech translation system. In: Proceedings of the International Workshop on Spoken Language Translation.

Barrows, H.S., 1987. Simulated (Standardized) Patients and other Human Simulations. Health Sciences Consortium, Chapel Hill, NC.

Belvin, R., May, W., Narayanan, S., Georgiou, P., Ganjavi, S., 2004. Creation of a doctor–patient dialogue corpus using standardized patients. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal.

Bernstein, A., Reinecke, K., 2007. Culturally adaptive software: moving beyond internationalization. In: Proceedings of the HCI International 2007.

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

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

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

Dybkjær, L., Bernsen, N.O., Minker, W., 2004. Evaluation and usability of multimodal spoken language dialogue systems. Speech Communication 43 (1–2), 33–54.

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

Ettelaie, Emil, Georgiou, Panayiotis G., Narayanan, Shrikanth, 2006. Cross-lingual dialog model for speech-to-speech translation. In: Proceedings of the International Conference on Spoken Language Processing, Pittsburgh, PA, September 2006.

Flanagan, J.L., 2004. Speech-centric multimodal interfaces. IEEE Signal Processing Magazine 21, 76–81.

Gao, Y., Gu, L., Zhou, B., Sarikaya, R., Afify, M., Kuo, H.K., et al., 2006. 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.

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

Jannach, D., Kreutler, G., 2005. 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.

Kao, C., Saleem, S., Prasad, R., Choi, F., Natarajan, P., Stallard, D., Krstovski, K., Kamali, M., 2008. Rapid development of an English/Farsi speech-to-speech translation system. In: Proceedings of IWSLT, Hawaii, Hawaii.

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

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

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

Larson, M.L., 1997. Meaning-Based Translation: A Guide to Cross-language Equivalence, 2nd ed. University Press of America.

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

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

MedicineNet, "Medicinenet.com," 2006.

Meeuwesen, L., Schaap, C., van der Staak, C., 1991. Verbal analysis of doctor–patient communication. Social Science and Medicine 32 (10), 1143–1150.

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

Onga, L.M.L., de Haesa, J.C.J.M., Hoosa, A.M., Lammesb, F.B., 1995. Doctor–patient communication: a review of the literature. Social Science and Medicine 40 (7), 903–918.

Oviatt, S., Coulston, R., Lunsford, R., 2004. When do we interact multimodally? Cognitive load and multimodal communication patterns. In: Proceedings of the Sixth International Conference on Multimodal Interfaces (ICMI), Pennsylvania, USA.

Oviatt, S., 2006. Human-centered design meets cognitive load theory: designing interfaces that help people think. In: Proceedings of the 14th Annual ACM International Conference on Multimedia.

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

Precoda, K., Zheng, J., Vergyri, D., Franco, H., Richey, C., Kathol, A., Kajarekar, S., 2007. IraqComm: a next generation translation system. In: Proceedings of the Interspeech, pp. 2841–2844.

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

Sanders, G.A., Bronsart, S., Condon, S., Schlenoff, C., 2008. Odds of successful transfer of low-level concepts: a key metric for bidirectional speech-to-speech machine translation in Darpa's transtac program. In: Proceedings of the LREC 2008.

Scheffler, K., Young, S., 2002. Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning. In: Proceedings of the Human Language Technology (HLT).

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

Shin, J., Georgiou, G., Narayanan, S., 2006. User modeling in a speech translation driven mediated interaction setting. In: Proceedings of the 1st ACM International Workshop on Human-centered Multimedia.

Walker, M.A., Rudnicky, A., Prasad, R., Aberdeen, J., et al., 2002. Darpa communicator: cross-system results for the 2001 evaluation. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP).

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

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