User Modeling in a Speech Translation Driven Mediated Interaction Setting

JongHo Shin
jonghosh@usc.edu

Panayiotis G. Georgiou
georgiou@sipi.usc.edu

Shrikanth Narayanan
shri@sipi.usc.edu

Viterbi School of Engineering,
University of Southern California.
3740 McClintock Ave., Room 400,
Los Angeles, CA 90089

ABSTRACT

The paper address user behavior modeling in a machine-mediated setting involving bidirectional speech translation. Specifically, usability data from doctor-patient dialogs involving a two way English-Persian speech translation system are analyzed to understand the nature, and extent, of user accommodation to machine errors. We consider user type – categorized along the classes of *Accommodating*, *Normal* and *Picky* – as it relates to the user's tendency to accept poor speech recognition and translation or retry to speak these again. For modeling, we employ a dynamic Bayesian network that can identify the user type with high accuracy after a few interactions of consistent user behavioral patterns. This model can be utilized for the design of machine strategies that can aid a user in operating the device more efficiently.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Miscellaneous—*User model*;

General Terms

Human Factors, Design

Keywords

User modeling, user-centered, user type, user interaction, speech-to-speech, translation, reasoning, inference, dynamic Bayesian network

1. INTRODUCTION

Spoken conversations have been recognized as the primary information delivery mechanism between humans and as speech and language technologies evolve, we can envision intelligent speech-enabled systems mediating dialogs between people, who do not share a language, through speech to speech translation. Significant progress has already been made in the creation of Human-Centered computer interfaces that aim in facilitating such an interaction by several research institutions, for example, [1, 8, 11, 15]. The goal of such systems is to be truly cognizant of the interaction, and serve as a communication aide than a mere message conduit. The existance of an effective Speech-to-Speech (S2S) translation system will have extreme impact in human communication, affecting healthcare, buisiness, social integration etc.

Drawing parallels with advances in human-machine spoken dialog systems, we can see that incorporating intelligence into a spoken language based mediating system requires, among other things, careful user modeling in conjunction with an effective dialog management. While there has been a fair amount of user modeling work in the context of human-machine spoken dialogs including user behavior simulation [3, 4], user goal and intention detection [5], specific user expertise modeling [6], evaluation techniques [7] and personalized dialog design [10], relatively little effort has been devoted in this regard to machine mediated human-human cross-lingual dialogs, the topic of this paper.

Modeling a user under problematic conditions can play an important role in designing flexible error handling mechanisms and in allowing for adaptation to the user, and hence improving not only the system performance but also the user satisfaction [6, 7]. The significance of modeling the user under problematic conditions in human-machine interaction is demonstrated for example by our prior work [14], where we highlighted the importance of repeating and rephrasing cues as key indicators. In the present work, we assume an asymmetric mediation scenario where one of the users has control over the device, and the dialog flow (such as a doctor in a doctor-patient interaction). We aim to model this user's behavior under problematic conditions based on information conveyed through the mediating channel (Human-Machine-Human). We, however, do not present analysis regarding the direct human-human channel, that contains cues of direct human-human interactions such as non verbal gestures, emotions, and mutual accommodation, but lacks verbal information transfer. In section 2, the system and data-set are described. In section 3, user behaviors observed in the dataset are analyzed, a dynamic Bayesian user model is in-

HCM'06, October 27, 2006, Santa Barbara, California, USA.

troduced and finally, the proposed model is evaluated. Discussion and conclusions are given in section 4.

2. SYSTEM AND DATA-SET

2.1 A Translation system with a push-to-talk interface

Transonics is a Speech-to-Speech translation device[1] that aims to enable two way spoken interactions between an English speaking doctor and a Persian (Farsi) speaking patient [8]. It is quite distinct from other machine speech recognition and translation projects as it is centered around human communication. In addition the approach we are taking in tackling this issue is unique as it is the first time the human variability is being considered as a factor affecting the successful outcome of an interaction.

This version of the system uses a push-to-talk speech activation modality. The push-to-talk interface minimizes recognition and translation errors by allowing users to verify translated concepts before executing the final decision of "speaking out" but has the disadvantage of creating less spontaneous and less natural interactions. The quality of the experience can be enhaced by adopting a user-centered interface design that allows for decision choices.

The domain of usage of the Transonics system is task-specific (or a goal-oriented) interactions between a doctor and a patient, rather than an open ended cross-lingual interaction between the two. In addition to high quality translation, we seek high user satisfaction, an important measurement metric in medical domain interactions [13].

To better understand the translation device operation, we can identify three distinct steps in a typical speech translation process. The first is the conversion from speech (audio) into a transcription of the spoken utterance (*Automatic Speech Recognition* or ASR), which is a lossy operation, i.e. often the transcript will not accurately represent what the user said. The second step is the translation. At this stage the text is mapped from the source language (English/Persian) to the target language (Persian/English). In our system, the user is given multiple translation options on a GUI: one possible translation tries to convey the concept exactly as the user said it, but is not guaranteed to achieve accurate translation; up to 4 other possible options convey the nearest concepts that the device knows how to translate accurately using utterance classification methods; a last option allows the user to reject the utterance and retry (either repeat or rephrase) again. The third stage is the conversion of the target language transcript from text to audio by synthesizing the speech using *Text-To-Speech* synthesis (TTS).

By design, the English-speaking doctor has exclusive control of the *Graphical User Interface* (GUI). Some of options provided to the doctor can be seen in Figure 1 and we can see in this example that the second option (labeled on the GUI: "I can try to translate this") is not exactly what the user said, but a better option since it is guaranteed to be translated correctly. As mentioned earlier, up to 4 options are presented to the user in this category.

[1]this consists of Automatic Speech Recognition, Translation, Dialog Management and Text-To-Speech synthesis as well as a visual output and control GUI.

Figure 1: Transonics system screen GUI. After speaking, the user(doctor) can choose one of several hypotheses presented on the GUI.

DARPA Evaluation results		
Off-line component metrics		
	English	Persian
ASR WER	11.5%	13.4%
	English to Persian	Persian to English
BLEU (text)	0.31	0.29
BLEU (ASR)	0.27	0.24
On-line system metrics (from 15 interactions)		
Overall concept transfer		78%

Table 1: DARPA evaluation on medical domain. Component and Concept measures as: ASR word error rate ,SMT BLEU score with the clean text transcript input or with the ASR output as an input.

The component level performance of Transonics – the word error rate, concept transfer rate and BLEU scores – are given in Table 1. These results stem from the evaluation under the DARPA Babylon program. The overall concept transfer rate (78%) denotes how many of the key concepts (such as medical symptom descriptions) were correctly transferred overall in both languages according to human observers for the 15 interactions examined in this paper. Also, in the Table 1 the word error rate (WER) and (IBM) BLEU [9] scores are provided.

2.2 Data-set

The data analyzed are from 15 interactions between doctors and actor patients. 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. Because of the dynamics created by the push-to-talk interface (controlled exclusively by the doctor), the doctor-side data contains abundant information we can utilize to model user behaviors in the mediated channel. The interactions were logged by the system and were also transcribed manually, hence making possible the automatic tagging of the *Retry/Accept* behavior and calculation of WER.

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 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.

Similar to the notion of expert/novice users, we consider the idea of accommodating and non-accommodating ("picky") user types under problematic situations. For example, our analysis indicates that for the same average WER, one user retried 95% of the time while another user retried only 65%. We have also observed that certain users are more accepting of minor errors in translation and recognition (e.g., insertion error resulting in "[And] do you have fever?" when they actually spoke "Do you have fever?") while others completely reject the machine's hypothesis as not their intended utterance, despite the fact that it conveys for all practical purposes the identical meaning. We therefore propose modeling the user controlling the device (doctor) in one of three categories (*Accommodating*, *Normal* and *Picky*), and we intend to create a model that can detect in which category the user belongs based on the user's action and speech recognition accuracy.

3.1 User types: Accommodating, Normal and Picky

For our off-line model, we cluster user types based on the total number of each user's retries, having observed minor differences in overall WER per user. We assume that accepting different ranges in WER as conceptualized on Figure 2 depends significantly on the user type and hence we define

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

Based on the 15 interactions analyzed in this work, we clustered the users using the k-means algorithm in the 3 classes as in Figure 3. 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 for *all* the utterances 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 the WER as a feature for the clustering of user types. Note that although the average WER is relatively constant from user to user, the error that users consider acceptable is not as demonstrated by the variable degree of retries.

Similar to categorizing the user types, we attempted to quantize the system performance as High-Quality (HQ) for low WER and Low-Quality (LQ) for high WER. This allows

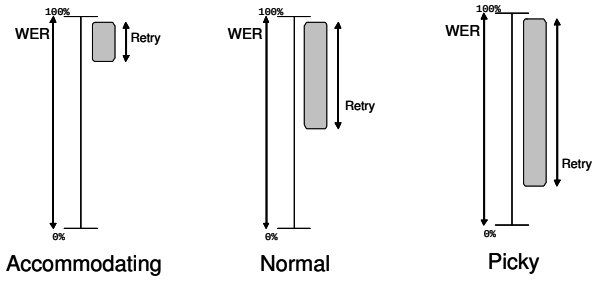

Figure 2: The Accommodating user "Retry" significantly less than the other users while the Picky significantly more. A user in between these extremes is defined to be a Normal user. WER is the Word Error Rate and the above graph demonstrates the ranges of WER for which each user type tends to "Retry"

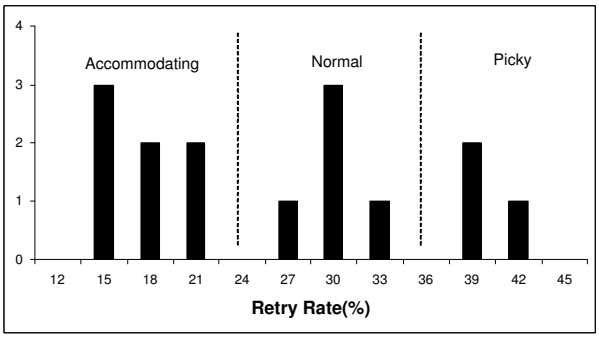

Figure 3: The quantized average retry rates over 15 interactions on the doctor side. The criteria (average "Retry" rate) based on the empirical analysis inspired us to categorize users into 3 types: Accommodating, normal, and picky.

for more robust modeling with limited data. By assuming that a user generally retries if the system performance falls below a threshold, we clustered the per-utterance WER into the 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. From Figure 4 we can see that there is significant difference in retries among user types when operating in the LQ region.

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 – reasoning under uncertainties. The network has been highlighted in the user modeling field across various applications[5, 2].

In this analysis the variables of user behavior (retry/accept)

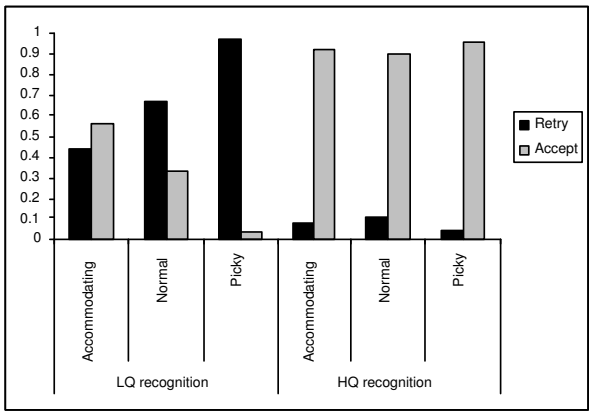

Figure 4: Conditional Probability Table(CPT) over user behaviors – "Retry" and "Accept". Each user type is represented with regard to Low Quality(LQ) and High Quality(HQ) performance.

Input:	User behavior("Retry" or "Accept") and HQ/LQ performance.
Output:	The most believable user type.
Initial:	User types with the same probability.
Step1:	Find the probability of each user type by a Bayesian reasoning.
Step2:	Update the prior of each user type.
Step3:	Check on confidence of user type probability.
Step4:	If not confident enough, go to "Step 1" otherwise, go to the next step. Return a user type with the highest belief.

Table 2: User type inference algorithm computes the probability of user types, Accommodating, Normal and Picky respectively. Each user type is predicted by a Bayesian reasoning and updated until one of them becomes believable.

and the feature of system confidence score (or for off-line processing WER) are the observed variables and the user type is an unknown variable. In Table 2, a user type is determined by applying Bayesian reasoning dynamically over time.

3.3 Training

We quantize the variables of user type (UT), behavior (B), and system accuracy (F)) where we chose $n = 3$ discrete levels for the user type, $m = 2$ for behavior and $k = 2$ for the WER.

To give a value for each discrete level, we can utilize a domain expert's knowledge or learn it from the data-set. A second method was adopted in this experiment where the values are learned from the training data-set in a cross-validation setting (using 14 out of 15 interactions for training and the remaining one for testing) allowing for presenting averaged results over a total of 15 experiments for the 15 interactions.

In constructing a Bayesian network, we decompose the

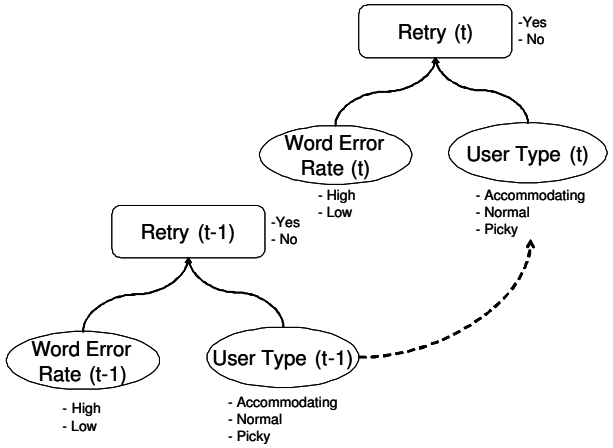

Figure 5: A dynamic Bayesian network is used to infer a user type over time in the mediated channel.

joint probability of behavior, user type and WER feature:

$$P(B, UT, F) = P(B|UT, F)P(UT)P(F) \quad (1)$$

where, B = user behavior, UT = user type, F = WER feature. It is assumed that there is no relationship between user type and feature.

During training, we employ the complete interaction to reason on the user type through the Maximum Likelihood Estimate (MLE) of $P(B|UT, F)$.

The transition probabilities are assumed to follow a parametric model that sets them higher within user type than across user types, and transition probabilities across the two opposite user types are lower than across neighboring user types.

3.3.1 A dynamic model

In a real time interaction, we need to observe a consistent behavior over time to strengthen our belief, but we also want to estimate the user type dynamically. We formulated this as a dynamic Bayesian network (DBN) shown in Figure 5. 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 the assumption.

3.4 Model Validation

We evaluated the automatic identification of the user type by employing the leave-one-out method, 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 and see relatively consistent user behavior patterns over time. Priors were set as equal (0.33) for the three user types. The success rate of classifying user type was 13 out of 15 by assuming a convergence of the DBN at the end of the available data. Both errors occurred in identifying the normal user type, and in both cases, 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, although we

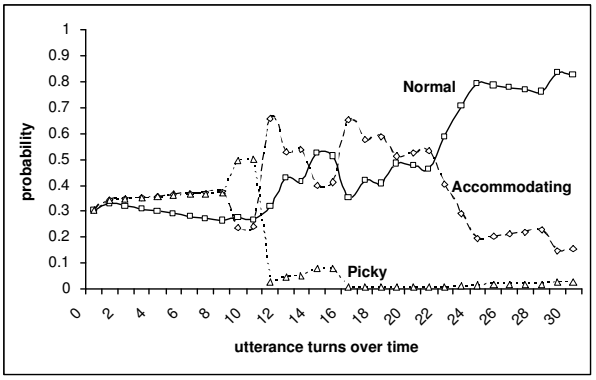

Figure 6: The belief that the user type is "Normal" is strengthened slowly over time.

have insufficient evidence to make a strong argument, that this may reflect a switching-user behavior depicting oscillating patterns.

Figure 6 shows one of the most challenging users to classify. The system in this case takes over 24 turns to eliminate the accommodating type, although it eliminated the Picky type from the 12th turn. A human analysis demonstrates 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.

4. DISCUSSION AND CONCLUSIONS

The paper addressed user behavior modeling in a machine-mediated setting involving bidirectional speech translation. Specifically, usability data from doctor-patient dialogs involving a two way English-Persian speech translation system was analyzed to understand two specific user behaviors under problematic conditions. Based on the observation of user behaviors, we simplified user types and quantized system performance that are effective for the small amount of data.

We considered the user type – quantized along the classes of *Accommodating*, *Normal* and *Picky* – as it relates to the user's tendency to tolerate system errors. The analysis employed a dynamic Bayesian network. We showed that we can identify the user type with high accuracy after a few interactions and consistent user behavioral patterns. This model can be utilized for the design of an efficient error handling mechanism; for example, in previous research [12], a correct interpretation of user's goal(intention) was helpful to deal with errors in human robot dialogs. We hope user modeling such as those attempted in our paper will enable future research in building systems that can appropriately adapt to users. We intend to employ more interaction data, to optimize for rapid identification of the user type, and we propose to investigate entropy measures in deciding the convergence point of the DBN.

5. ACKNOWLEDGEMENTS

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

6. REFERENCES

[1] 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.

[2] 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.

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

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

[5] 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.

[6] 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.

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

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

[9] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.

[10] A. N. Pargellis, J. Kuo, and C. H. Lee. An automatic dialogue generation platform for personalized dialogue applications. *Speech Communication*, 42(3-4):329–351, 2004.

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

[12] 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.

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

[14] J. H. Shin, S. Narayanan, L. Gerber, A. Kazemzadeh, and D. Byrd. Analysis of user behavior under error conditions in spoken dialogs. In *Proc. of 5th ICSLP*, 2002.

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