Analyzing the Multimodal Behaviors of Users of a Speech-to-Speech Translation Device by using Concept Matching Scores

JongHo Shin, Panayiotis G. Georgiou, and Shrikanth Narayanan
Viterbi School of Engineering
Speech Analysis and Interpretation Laboratory
University of Southern California
Los Angeles, California 90089
Email: jonghosh@usc.edu, georgiou@sipi.usc.edu and shri@sipi.usc.edu

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

I. Introduction

Current speech translation technologies support real time spoken language translation, and are applicable in many areas such as medical services, security and disaster relief, and business communication. Successful prototype examples include Verbmobil [1], a system for multilingual scheduling, including airline and hotel reservations; Transonics [2], a communication aid for use by doctors; and MASTOR [3], a multilingual automatic *Speech-to-Speech* (S2S) translation system for a variety of domains.

Although there has been intensive research on speech recognition technology [4] and on machine translation [5], studies modeling users of S2S translation systems have rarely been conducted. The usefulness of studying user behaviors has already been demonstrated for automated spoken dialog systems and it is our hypothesis that similar benefits from user studies can be achieved in S2S systems. For our work we draw motivation from existing studies such as on system evaluation and multimodal interfaces. Kamm and Walker [6] measured the performance of a spoken dialog system in terms of task success and cost (number of turns). These two factors were utilized for maximizing user satisfaction. Oviatt et al [7] reported that users of a spoken dialog system tend to employ more of the available modalities as cognitive load increases with intensifying task difficulty and communicative complexity. Also, Foster et al [8], in analyzing text prediction models, have argued that by modeling the user they can provide improved text prediction. For S2S translation systems, user studies can cover a vast range of topics, such as language, culture, environment, education, and belief systems [9].

In particular, it is important to study the quality of the concepts transferred through the S2S translation systems, and the level of transfer errors users are willing to accept when using such a system. For instance, a significant part of the machine error stems from speech recognition and users will be able to gauge in some part the degree of degradation if they can observe the transcribed text of what they said. Experimental evidence has shown that some users are more accommodating to these system errors and still go ahead and accept erroneous speech recognizer output as acceptable for translation, knowing well that they increase the chance of poor concept transfer.

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

Like a human translator, a translation device transfers meaning from one language, such as English, to another language, such as Farsi (Persian) [9]. The process is inherently lossy. Vocabulary words and phrases need to be changed to their closest representation in the target language, but will often be remapped to more distant equivalents, and grammar and syntax will also degrade. As a result, the original meaning will be altered at several different levels [10], conveying it sometimes quite closely and sometimes poorly. It is important to measure how well meaning is transferred by translation devices. The

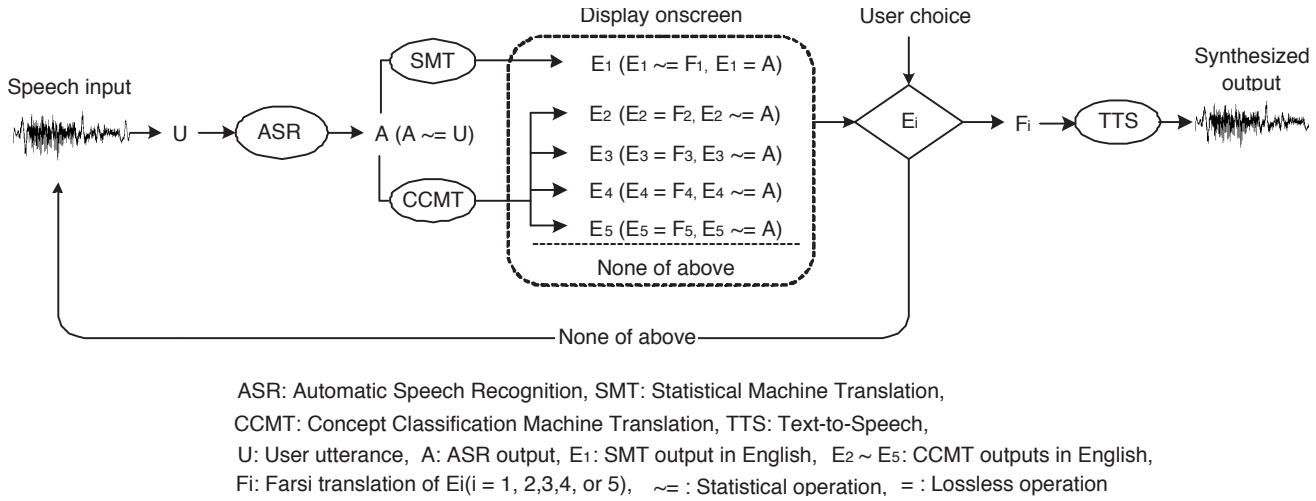

Fig. 1. The internal procedure of generating speech translation candidates implemented in the Transonics system. A doctor uses two-modality interface (push-to-talk), and sees up to five candidates onscreen; one Machine Translation (MT) candidate (E_1), and up to four Classifier candidates (E_2,E_3,E_4,E_5).

existing *text translation* metrics, such as BLEU [11] and NIST [12] scores are based on comparisons of several human translations with system-produced translations using n-gram matching. In the study of this paper, to measure how well meaning is transferred by S2S translation system, we introduce a measure called concept matching score. This score refers to the number of concepts in a user utterance in the source language that is carried over to the machine-produced utterance in the target language. We evaluate the performance of the S2S system and its subcomponents in terms of the concept agreement between the input and output according to human annotators.

The paper is organized as follows. Section II describes the S2S system used for the experiments and section III, the data collection and annotation. We present results on multimodal versus single modality usage in Section IV-A, statistics on user error tolerance in IV-B and analysis on the quality of user choices in IV-C. Discussion and conclusions follow in sections V and VI.

II. The Transonics System

Transonics [2] is a speech-to-speech (S2S) translation system, which facilitates two way spoken interactions between English-speaking doctors, and Farsi-speaking patients. This system is aimed at task-oriented interactions in the medical domain.

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

In simple terms, the Transonics design processes the input speech as follows: First, it converts the speech into text (Automatic Speech Recognition – ASR); second, it converts the text into the target language (Machine Translation – MT); third, it plays out the translated text (Text-To-Speech synthesis TTS). The MT step operates in one of two modes resulting in multiple translation alternatives: The phrase-based translation (often called Statistical Machine Translation – SMT); and the concept based translation (Concept Classification – CCMT). The English speaker sees the various options on the screen after the MT step. We always show one option (E_1) that can be transferred through the SMT path, and up to 4 options ($E_2 - E_5$) that can be transferred through the CCMT path. The CCMT path has the advantage that it provides a very accurate back translation since the concepts known by the CCMT were previously humanly translated. Thus options $E_2 - E_5$ will be

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

transferred very accurately in the target language, while option E_1 will undergo some further channel loss.

Figure 1 graphically shows the above description and defines the symbols for subsequent clarity. In short: U is the original user input; A is the ASR belief ($A \simeq U$); $E_1 = A$ is the text that will be translated through the SMT and generate (lossy operation) F_1 ($F_1 \simeq E_1 \simeq U$); and $E_2 - E_5$ is the text already translated and mapped back ("non-lossy", human mapping) into English through CCMT ($U \simeq A \simeq F_i = E_i,\ \forall i = \{2, 3, 4, 5\}$).

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

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

III. Methods

A. Data collection

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

The *Concept Matching Score* (CMS) was based on the Linguistic Data Consortium's human assessment metrics [13]. Ma and Cieri [13] observe "Adequacy refers to the degree to which the translation communicates information present in the original or in the best of breed translation that serves as a proxy to the original." Based on that the concept matching score compares the number of concepts in an original utterance (source) and the target utterance (destination), either through translation or within the same language, e.g. through a lossy speech recognition channel. The score guidelines for CMS are:

1.0: All concepts are transferred.
0.8: Most concepts are transferred.
0.6: Many concepts are transferred.
0.4: Some concepts are transferred, such that users may sometimes get the whole meaning.
0.2: Few concepts are transferred, such that users rarely get the whole meaning.
0.0: None of the concepts are transferred.

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

Example, Average CMS=0.8:
User (U): DO YOU HAVE DOUBLE VISION
Translation (F_1): VyA SmA v dyd dvgAnh dAryd
(DOES AND YOU HAVE A DOUBLE VISION)

IV. Results and Analysis

In the study of this paper, we attempt to address three hypotheses. The methods and results are presented in the following three sections IV-A, IV-B, and IV-C.

- **Hypothesis 1:** A multimodal interface, employing both the audio and text modalities, will be better than a single-modality interface utilizing audio only, in terms of translation quality.
- **Hypothesis 2:** Users will accept certain errors from the utterances provided by the system. The degree of degradation in terms of concept representation that different users are willing to accept varies.
- **Hypothesis 3:** Improving the feedback as to when it is appropriate to employ the CCMT path can yield improved translation quality.

A. Multimodal versus single-modality interface

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

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

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

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

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

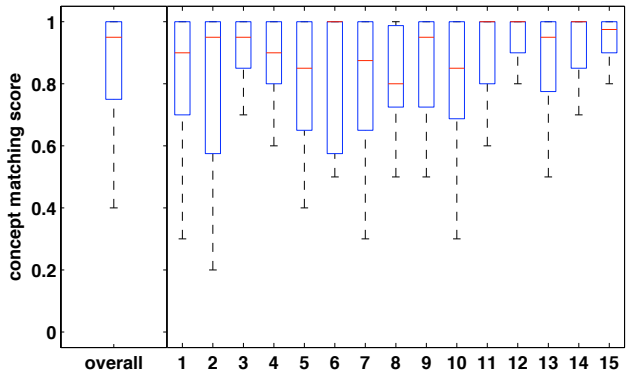

Fig. 3. A box-plot and concept matching scores of user accepted utterances in the 15 sets of interactions. User retry utterances were not included in the total.

B. User Error Tolerance

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

The left part of Figure 3 shows a box-plot of concept matching scores of user selected utterances in the 15 sets of interactions. The median score of 0.95 indicates that more than half of the users accepted the onscreen machine-produced utterances when they contained 95% of the concepts in the original utterances. The standard deviation (unbiased) was 0.21, and the mean absolute deviation was 0.17. We also note that users accepted machine-produced utterances with concept matching scores as low as 0.4, which reveals quite accommodating users. The mean concept matching score was 0.84, indicating that users on average are accepting of 16% concept loss from the speech recongizer.

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

C. The quality of user choices

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

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

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

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

In the above analysis and in figure 4 we considered the input of the user as the detected path. When analyzing the SMT path and the CCMT path for the same utterance however, we found that in 13% of the data the users decided to take through the SMT path, performance would have been improved if the users had chosen the CCMT path. In this 13% of the data, we found that the average concept matching score of 0.83 was degraded to 0.62 through the SMT path, but the average concept matching score through the CCMT was 0.75.

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

$$\text{CMS}[U, E_1] > \text{CMS}[U, E_i] \quad \forall i = \{2, 3, 4, 5\} \tag{1}$$

but,

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

Our hypothesis, that improving dynamic user feedback on when it is appropriate to chose the CCMT path, can yield translation improvements has proven true. To optimize performance, the translation system needs strategies to elicit better

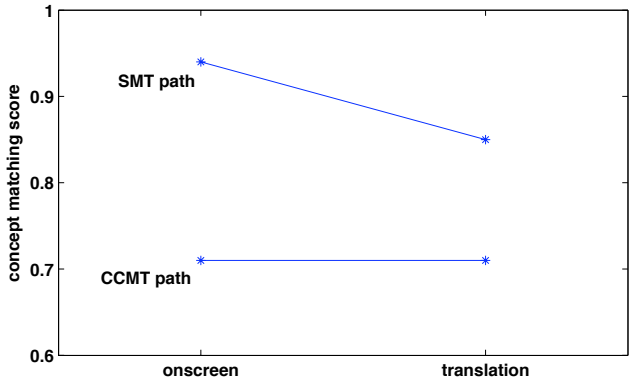

Fig. 4. Concept matching scores of the onscreen utterances selected by users with the push-to-talk interface of the Transonics, and concept matching scores of the corresponding final translations. User-selected utterances were processed through the SMT path and the CCMT path. SMT is a statistical machine translation and CCMT is concept classification machine translation.

user choices onscreen in cases of discrepant translation quality. This requires better self-assessment by the system as to its expected translation accuracy through the SMT and better guidance to the user on the expected degradation. To this end, we introduced back-translation functionality in the SMT path in the newer version of the system. Although the effect of this will be studied in future work, preliminary observations seem to indicate that this discourages users significantly more than it was meant to, and encourages complete rejection.

V. Discussion

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

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

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

VI. Conclusion

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

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

VII. Acknowledgement

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

References

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

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

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

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

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

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

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

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

[9] J. Cohen, "From meaning to meaning: the influence of translation techniques on non-english focus group research," *Qualitative health research*, vol. 11, no. 4, pp. 568–579, 2001.

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

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

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

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