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