Transonics: A Practical Speech-to-Speech Translator for English-Farsi Medical Dialogues

Emil Ettelaie, Sudeep Gandhe, Panayiotis Georgiou, Kevin Knight, Daniel Marcu, Shrikanth Narayanan , David Traum
University of Southern California
Los Angeles, CA 90089
ettelaie@isi.edu, gandhe@ict.usc.edu, georgiou@sipi.usc.edu, knight@isi.edu, marcu@isi.edu, shri@sipi.usc.edu, traum@ict.use.edu

Robert Belvin
HRL Laboratories, LLC
3011 Malibu Canyon Rd.
Malibu, CA 90265
rsbelvin@hrl.com

Abstract

We briefly describe a two-way speech-to-speech English-Farsi translation system prototype developed for use in doctor-patient interactions. The overarching philosophy of the developers has been to create a system that enables effective communication, rather than focusing on maximizing component-level performance. The discussion focuses on the general approach and evaluation of the system by an independent government evaluation team.

1 Introduction

In this paper we give a brief description of a two-way speech-to-speech translation system, which was created under a collaborative effort between three organizations within USC (the Speech Analysis and Interpretation Lab of the Electrical Engineering department, the Information Sciences Institute, and the Institute for Creative Technologies) and the Information Sciences Lab of HRL Laboratories. The system is intended to provide a means of enabling communication between monolingual English speakers and monolingual Farsi (Persian) speakers. The system is targeted at a domain which may be roughly characterized as "urgent care" medical interactions, where the English speaker is a medical professional and the Farsi speaker is the patient. In addition to providing a brief description of the system (and pointers to papers which contain more detailed information), we give an overview of the major system evaluation activities.

2 General Design of the system

Our system is comprised of seven speech and language processing components, as shown in Fig. 1. Modules communicate using a centralized message-passing system. The individual subsystems are the Automatic Speech Recognition (ASR) subsystem, which uses n-gram Language Models (LM) and produces n-best lists/lattices along with the decoding confidence scores. The output of the ASR is sent to the Dialog Manager (DM), which displays the n-best and passes one hypothesis on to the translation modules, according to a user-configurable state. The DM sends translation requests to the Machine Translation (MT) unit. The MT unit works in two modes: Classifier based MT and a fully Stochastic MT. Depending on the dialogue manager mode, translations can be sent to the unit selection based Text-To-Speech synthesizer (TTS), to provide the spoken output. The same basic pipeline works in both directions: English ASR, English-Persian MT, Persian TTS, or Persian ASR, Persian-English MT, English TTS.

There is, however, an asymmetry in the dialogue management and control, given the desire for the English-speaking doctor to be in control of the device and the primary "director" of the dialog.

The English ASR used the University of Colorado *Sonic* recognizer, augmented primarily with LM data collected from multiple sources, including

Proceedings of the ACL Interactive Poster and Demonstration Sessions,
pages 89–92, Ann Arbor, June 2005.

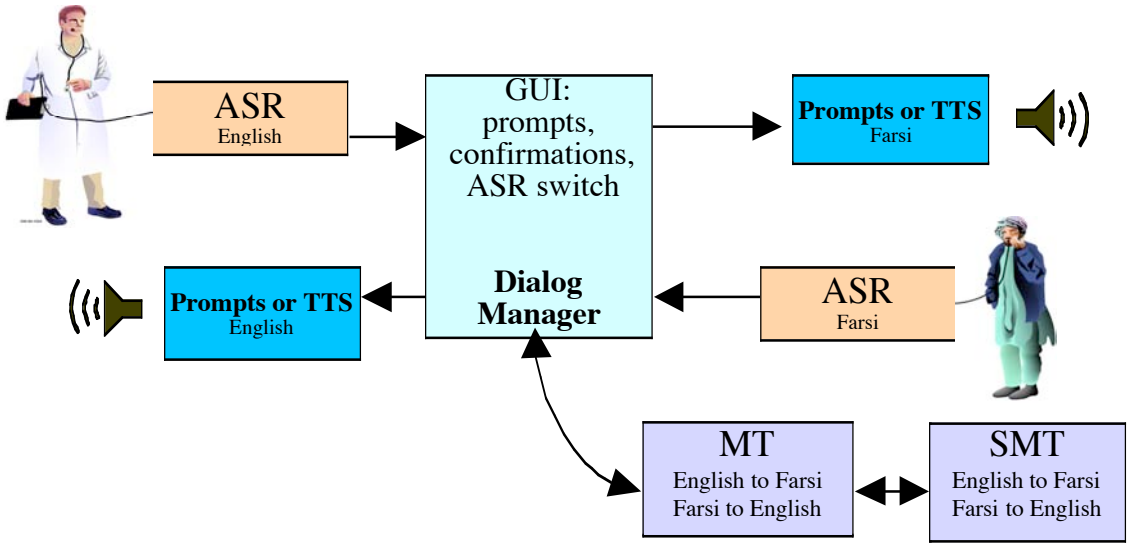

Figure 1: Architecture of the Transonics system. The Dialogue Manager acts as the hub through which the individual components interact.

our own large-scale simulated doctor-patient dialogue corpus based on recordings of medical students examining standardized patients (details in Belvin et al. 2004).[1] The Farsi acoustic models r equired an eclectic approach due to the lack of existing labeled speech corpora. The approach included borrowing acoustic data from English by means of developing a sub-phonetic mapping between the two languages, as detailed in (Srinivasamurthy & Narayanan 2003), as well as use of a small existing Farsi speech corpus (FARSDAT), and our own team-internally generated acoustic data. Language modeling data was also obtained from multiple sources. The Defense Language Institute translated approximately 600,000 words of English medical dialogue data (including our standardized patient data mentioned above), and in addition, we were able to obtain usable Farsi text from mining the web for electronic news sources. Other smaller amounts of training data were ob tained from various sources, as detailed in (Narayanan et al. 2003, 2004). Additional detail on development methods for all of these components, system integration and evaluation can also be found in the papers just cited.

The MT components, as noted, consist of both a Classifier and a stochastic translation engine, both developed by USC-ISI team members. The English Classifier uses approximately 1400 classes consisting mostly of standard questions used by medical care providers in medical interviews. Each class has a large number of paraphrases associated with it, such that if the care provider speaks one of those phrases, the system will identify it with the class and translate it to Farsi via table-lookup. If the Classifier cannot succeed in finding a match exceeding a confidence threshold, the stochastic MT engine will be employed. The stochastic MT engine relies on n-gram correspondences between the source and target languages. As with ASR, the performance of the component is highly dependent on very large amounts of training data. Again, there were multiple sources of training data used, the most significant being the data generated by our own team's English collection effort, supported by translation into Farsi by DLI. Further details of the MT components can be found in Narayanan et al., *op.cit*.

3 Enabling Effective Communication

The approach taken in the development of Transonics was what can be referred to as the *total communication pathway*. We are not so concerned with trying to maximize the performance of a given component of the system, but rather with the effectiveness of the system as a whole in facilitating actual communication. To this end, our design and development included the following:

[1] *Standardized Patients* are typically actors who have been trained by doctors or nurses to portray symptoms of particular illnesses or injuries. They are used extensively in medical education so that doctors in training don't have to "practice" on real patients.

i. an "educated guess" capability (system guessing at the meaning of an utterance) from the Classifier translation mechanism—this proved very useful for noisy ASR output, especially for the restricted domain of medical interviews.

ii. a flexible and robust SMT good for filling in where the more accurate Classifier misses.

iii. exploitation of a partial n-best list as part of the GUI used by the doctor/medic for the English ASR component and the Farsi-to-English translation component.

iv. a dialog manager which in essence occasionally makes "suggestions" (for next questions for the doctor to ask) based on query sets which are topically related to the query the system believes it recognized the doctor to have spoken.

Overall, the system achieves a respectable level of performance in terms of allowing users to follow a conversational thread in a fairly coherent way, despite the presence of frequent ungrammatical or awkward translations (i.e. despite what we might call *non-catastrophic* errors).

4 Testing and Evaluation

In addition to our own laboratory tests, the system was evaluated by MITRE as part of the DARPA program. There were two parts to the MITRE evaluations, a "live" part, designed primarily to evaluate the overall task-oriented effectiveness of the systems, and a "canned" part, designed primarily to evaluate individual components of the systems.

The live evaluation consisted of six medical professionals (doctors, corpsmen and physician's assistants from the Naval Medical Center at Quantico, and a nurse from a civilian institution) conducting unrehearsed "focused history and physical exam" style interactions with Farsi speakers playing the role of patients, where the English-speaking doctor and the Farsi-speaking patient communicated by means of the Transonics system. Since the cases were common enough to be within the realm of general internal medicine, there was no attempt to align ailments with medical specializations among the medical professionals.

MITRE endeavored to find primarily monolingual Farsi speakers to play the role of patient, so as to provide a true test of the system to enable communication between people who would otherwise have no way to communicate. This goal was only partially realized, since one of the two Farsi patient role-players was partially competent in English.[2] The Farsi-speaking role-players were trained by a medical education specialist in how to simulate symptoms of someone with particular injuries or illnesses. Each Farsi-speaking patient role-player received approximately 30 minutes of training for any given illness or injury. The approach was *similar* to that used in training standardized patients, mentioned above (footnote 1) in connection with generation of the dialogue corpus.

MITRE established a number of their own metrics for measuring the success of the systems, as well as using previously established metrics. A full discussion of these metrics and the results obtained for the Transonics system is beyond the scope of this paper, though we will note that one of the most important of these was task-completion. There were 5 significant facts (5 distinct facts for each of 12 different scenarios) that the medical professional should have discovered in the process of interviewing/examining each Farsi patient. The USC/HRL system averaged 3 out of the 5 facts, which was a slightly above-average score among the 4 systems evaluated. A "significant fact" consisted of determining a fact which was critical for diagnosis, such as the fact that the patient had been injured in a fall down a stairway, the fact that the patient was experiencing blurred vision, and so on. Significant facts did not include items such as a patient's age or marital status.[3] We report on this measure in that it is perhaps the single most important component in the assessment, in our opinion, in that it is an indication of many aspects of the system, including *both* directions of the translation system. That is, the doctor will very likely conclude correct findings only if his/her question is translated correctly to the patient, and also if the patient's answer is translated correctly for the doctor. In a true medical exam, the doctor may have

[2] There were additional difficulties encountered as well, having to do with one of the role-players not adequately grasping the goal of role-playing. This experience highlighted the many challenges inherent in simulating domain-specific spontaneous dialogue.

[3] Unfortunately, there was no baseline evaluation this could be compared to, such as assessing whether any of the critical facts could be determined without the use of the system at all.

other means of determining some critical facts even in the absence of verbal communication, but in the role-playing scenario described, this is very unlikely. Although this measure is admittedly coarse-grained, it simultaneously shows, in a crude sense, that the USC/HRL system compared favorably against the other 3 systems in the evaluation, and also that there is still significant room for improvement in the state of the art.

As noted, MITRE devised a *component* evaluation process also consisting of running 5 scripted dialogs through the systems and then measuring ASR and MT performance. The two primary component measures were a version of BLEU for the MT component (modified slightly to handle the much shorter sentences typical of this kind of dialog) and a standard Word-Error Rate for the ASR output. These scores are shown below.

Table 1: Farsi BLEU Scores

	IBM BLEU ASR	IBM BLEU TEXT
English to Farsi	0.2664	0.3059
Farsi to English	0.2402	0.2935

The reason for the two different BLEU scores is that one was calculated based on the ASR component output being translated to the other language, while the other was calculated from human transcribed text being translated to the other language.

Table 2: HRL/USC WER for Farsi and English

	English	Farsi
WER	11.5%	13.4%

5 Conclusion

In this paper we have given an overview of the design, implementation and evaluation of the Transonics speech-to-speech translation system for narrow domain two-way translation. Although there are still many significant hurdles to be overcome before this kind of technology can be called truly robust, with appropriate training and two cooperative interlocutors, we can now see some degree of genuine communication being enabled. And this is very encouraging indeed.

6 Acknowledgements

This work was supported primarily by the DARPA CAST/Babylon program, contract N66001-02-C-6023.

References

R. Belvin, W. May, S. Narayanan, P. Georgiou, S. Ganjavi. 2004. Creation of a Doctor-Patient Dialogue Corpus Using Standardized Patients. In Proceedings of the *Language Resources and Evaluation Conference* (LREC), Lisbon, Portugal.

S. Ganjavi, P. G. Georgiou, and S. Narayanan. 2003. Ascii based transcription schemes for languages with the Arabic script: The case of Persian. In *Proc. IEEE ASRU*, St. Thomas, U.S. Virgin Islands.

S. Narayanan, S. Ananthakrishnan, R. Belvin, E. Ettelaie, S. Ganjavi, P. Georgiou, C. Hein, S. Kadambe, K. Knight, D. Marcu, H. Neely, N. Srinivasamurthy, D. Traum and D. Wang. 2003. Transonics: A speech to speech system for English-Persian Interactions, *Proc. IEEE ASRU*, St. Thomas, U.S. Virgin Islands.

S. Narayanan, S. Ananthakrishnan, R. Belvin, E. Ettelaie, S. Gandhe, S. Ganjavi, P. G. Georgiou, C. M. Hein, S. Kadambe, K. Knight, D. Marcu, H. E. Neely, N. Srinivasamurthy, D. Traum, and D. Wang. 2004. The Transonics Spoken Dialogue Translator: An aid for English-Persian Doctor-Patient interviews, in *Working Notes of the AAAI Fall symposium on Dialogue Systems for Health Communication*, pp 97--103.

N. Srinivasamurthy, and S. Narayanan. 2003. Language adaptive Persian speech recognition. In proceedings of *Eurospeech* 2003.