Limited-Domain Speech-to-Speech Translation between English and Pashto

Kristin Precoda, Horacio Franco, Ascander Dost, Michael Frandsen, John Fry, Andreas Kathol, Colleen Richey, Susanne Riehemann, Dimitra Vergyri, Jing Zheng
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025
Christopher Culy
FX Palo Alto Laboratory, Inc.
3400 Hillview Ave., Building 4
Palo Alto, CA 94304

Abstract

This paper describes a prototype system for near-real-time spontaneous, bidirectional translation between spoken English and Pashto, a language presenting many technological challenges because of its lack of resources, including both data and expert knowledge. Development of the prototype is ongoing, and we propose to demonstrate a fully functional version which shows the basic capabilities, though not yet their final depth and breadth.

1 Introduction

This demonstration will present a prototype system for bidirectional speech-to-speech translation within a limited semantic domain, that of first encounters between a patient and a medical professional. A major goal of the work is to explore techniques that are appropriate for languages that are not of great commercial interest and that consequently are lacking in data and resources of many kinds. The system has been developed for American English and Pashto, one of the major languages of Afghanistan, which presents a variety of challenges for both data-intensive and knowledge-based approaches.

It should be noted that the system must be viewed as one component in a real-time transaction between two cooperating humans. The ultimate goal of those humans is to exchange information by whatever means is effective: not, necessarily, to rely exclusively on the system's output, but to use it in combination with other, non-speech modalities of conveying meaning and with ordinary world knowledge.

The final system is intended to run on a handheld device, such as a PDA, with its attendant memory and speed limitations and restriction to integer-only computation. Most components of the demonstration system run in a PocketPC emulation environment on a Windows laptop, with a few in the full Windows environment. All aspects of the prototype system are undergoing active development.

2 Overall Architecture

A simple description of the architecture is as follows. The system is controlled by the English speaker, who is expected to have greater technological familiarity and who has the benefit of visual feedback on the system performance. Spoken input, in either English or Pashto, is recognized by SRI's small-footprint DynaSpeak® recognizer, and an ordered list of hypotheses is produced. The most likely hypothesis is input to SRI's Gemini natural language parser/generator (Dowding et al. 1993), which attempts to parse the speech recognition output. Handling of possible errors or failures will be discussed in Section 3.

When a successful parse is obtained, Gemini creates a quasi-logical form representing the meaning of the sentence. In general, multiple quasi-logical forms may be created, reflecting differing interpretations of the input sentence. These forms, which are domain

independent and serve here as an interlingua, can be ordered by heuristically assigning preferences or dispreferences to the parsing rules applied to create them. Gemini uses a grammar for the target language to generate a translation from the most-preferred interpretation possible, and outputs a textual representation of the translation.

Theta, a small-footprint concatenative synthesizer from Cepstral LLC (Cepstral LLC 2004), then produces synthetic spoken output in the target language from the textual representation generated by Gemini. The Pashto voice was created specifically for this project.

An English and a Pashto version of each component are called by a single application which includes a graphical user interface. A screen shot of the interface is shown in Figure 1.

Figure 1. Screen shot of prototype system interface.

3 Redundancy and Handling Infelicities

As in any complex system, performance can differ from the ideal in a number of ways, and it is important for the system to provide alternative ways to support successful communication. Some kinds of subideal performance and recovery approaches are described here.

The most likely hypothesis output by the speech recognizer may not be exactly what was spoken. When the input speech is English, the English speaker can see whether the most likely hypothesis (shown in the "Input" box in Figure 1) is correct or approximately correct. If the English speaker judges the hypothesis to not be close enough to the intended text, s/he may either repeat the utterance or click on the "Guesses…" button to see an ordered list of the best hypotheses. A sample list is shown in Figure 2. If the correct text is on this list, the user may select it to submit for translation. When the input speech is in Pashto, this functionality is less useful, as the Pashto speaker is not assumed to be able to read, even if the hypotheses were displayed in Pashto script.

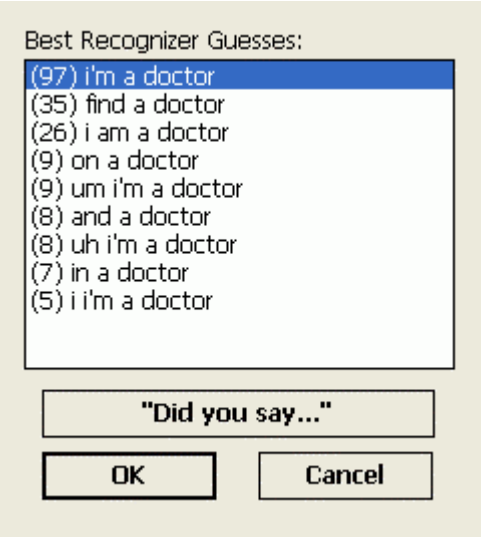

Figure 2. Sample ordered list of recognizer hypotheses.

Once a recognition hypothesis has been submitted for translation, several possible problems can arise. Pashto is a moderately inflected, split-ergative Indo-European language, and for Pashto in particular, recognition errors may lead to apparent lack of syntactic agreement between elements of the sentence which should (and did in fact) agree. As Gemini tries to generate a full parse of the input, it has the option of using parse rules that relax agreement requirements. These rules are dispreferred and a parse built upon them may be kept only if a full, grammatically correct parse cannot be completed. Another possible problem is that unknown words, some grammatical constructions, and input errors may render any full parse unachievable. In this case, fallback strategies can be applied to translate partial parses, fragments, or any known words. Other strategies are currently in development.

Another class of approaches for assisting communication allows the English speaker to quickly perform certain actions or play high-frequency phrases to the Pashto speaker. If there is background noise or distractions or the TTS quality is not high enough for easy comprehension, pressing the "Replay" button will immediately play back the last translation result. "Ask for Rep" plays a prerecorded sentence asking the Pashto speaker to repeat what s/he said. Several other useful phrases are available by clicking on the "Phrases…" button and then selecting from the screen shown in Figure 3.

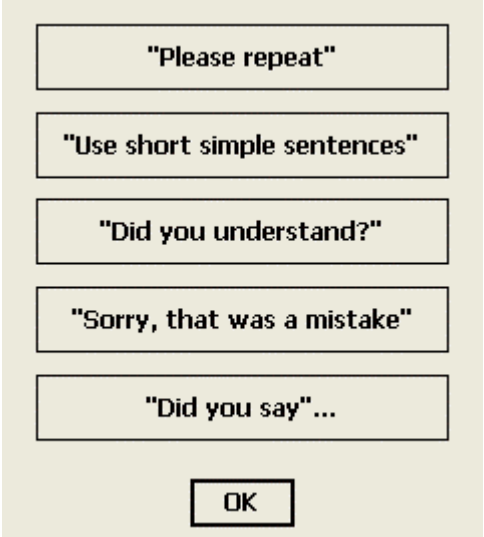

Figure 3. Prerecorded Pashto phrases which can be played back with a single click.

4 Sample Interaction

The table below shows an excerpt of a dialog between an English and a Pashto speaker, both new to this system, whose interaction was part of an informal trial run by the MITRE Corporation in February 2004 for the DARPA CAST program. The English speaker, a doctor, had just received eighteen minutes of training in how to use the system and had had no other exposure to it. The Pashto speaker, playing the role of an injured patient, had received training in complaints consistent with a particular injury scenario and had seen others use the system, but had not interacted with the system himself. Except as noted, the translations are understandable.

Spoken input	System result
I am a doctor, can I help you?	[failure to translate because two sentences]
I am a doctor	z@ ddAkttar y@m
Can I help you?	AyA z@ d@ tA sara komak kaw@lay S@m
ho, mehrabAni w@krra	yes make [partial translation; full translation should have been "yes please do"]
Where are you hurt?	[rerecorded after one misrecognition] t@ cherta khugx ye
ghwagx aw ugxa	[misrecognized repeatedly; unable to give meaningful translation; should have been "ear and shoulder"]
Can you breathe?	AyA t@ sA akhIst@lay Se
na, mUSkel lar@m	no I have the problem
Do you have pain?	AyA t@ dard lare
zAyt	of much [minor misrecognition; translation should have been "much"]
Do you take medications?	AyA t@ dawAuna akhle [incorrect plural form of dawA, but understood by Pashto speaker anyway]
na	no
Do you want pain medication?	AyA t@ d@ dard dawA ghwArre
ho	yes
Do you have allergies?	AyA t@ hasAsiyatuna lare
ho	yes
What are your allergies?	stA hasAsiyat ts@ day
antibyutik	[misrecognized but correctly understood by doctor as "antibiotics"]

5 Challenges

Three main challenges face this project. First, commercially nonviable languages, such as Pashto, often offer very few linguistic resources (such as linguistic descriptions, acoustic data, texts, language processing tools). The lack of resources makes development more difficult, and severely constrains the approaches that are viable: approaches that rely on large corpora cannot be used. In addition, there is a shortage of literate speakers who are available to act as consultants, and a scarcity of basic knowledge about the language. This impedes progress and renders difficult approaches that rely on a large body of hand-crafted translation rules. The challenge of having no single person who has a deep understanding of both the language and the technology and who can serve as a bridge between them is substantial, and causes more iterative development than is ordinarily the case when bilingual technologists are available, as newly discovered phenomena or new understanding cause revisions of previous work.

A second major challenge is due to the nature of the domain and the underlying concept of operations. Real speech occurs in noisy environments, has disfluencies, and is highly variable (e.g., phrasings, dialect differences). In addition, the output of a speech recognizer contains more and qualitatively different errors than typical text input to automatic translation systems. While the problems of real speech are not unique to this project, they are magnified by the fact that the non-English speakers will largely be unsophisticated users of technology, who will often be using the system for the first and only time. The system

must work well from the very first utterance – there cannot be much of a learning curve. This applies to the translation quality and to other aspects of the system, such as the synthetic speech, as speakers are not familiar with synthesized Pashto speech. These speakers are also not expected to be literate, and their understanding will not be bolstered by the extra redundancy and capabilities that the display offers to the English speaker.

A third major challenge is posed by the handheld device platform with its computational and storage limitations, and the near-real-time requirement of the envisioned usage. The restriction to integer-only computation is most serious for the speech recognition, as nearly all medium- or large-vocabulary speech recognizers perform extensive floating-point computations, and the conversion of a speech recognizer to use only integer computation required considerable effort. The severe memory limitations perhaps impact most the parsing/generation components of the system.

6 Summary

We have described a prototype of a spoken language translation system for use by English-speaking medical personnel treating Pashto-speaking patients. The system, which is targeted to a handheld device, is being developed with extremely little Pashto-language data of any kind. It builds on medium-vocabulary speaker-independent HMM-based speech recognition, rule-based translation and supporting methods, and concatenative synthesis. While the system is certainly still under development, it has provided a reasonable proof of concept in informal trials.

Acknowledgments

This work was supported under SPAWAR contract N66001-99-D-8504 with funding from DARPA. Many thanks are due to Mohammad Shahabuddin Khan, David Kale, Robert J. Podesva, Valerie Wagner, and many Pashto speakers who worked with us in various capacities.

References

Cepstral, LLC. 2004. *Theta: Small footprint text-to-speech synthesizer*, Pittsburgh, PA, http:// www.cepstral.com

J. Dowding, J. M. Gawron, D. E. Appelt, J. Bear, L. Cherny, R. Moore, and D. B. Moran. 1993. Gemini: A natural language system for spoken language understanding. *Proceedings of the Thirty-First Annual Meeting of the Association for Computational Linguistics*, 54-61.