A Wearable Headset Speech-to-Speech Translation System

Kristе Krstovski, Michael Decerbo, Rohit Prasad, David Stallard, Shirin Saleem, Premkumar Natarajan
Speech and Language Processing Department
BBN Technologies
10 Moulton Street, Cambridge, MA, 02138
{krstovski, mdecerbo, rprasad, stallard, ssaleem, prem}@bbn.com

Abstract

In this paper we present a wearable, headset integrated eyes- and hands-free speech-to-speech (S2S) translation system. The S2S system described here is configured for translingual communication between English and colloquial Iraqi Arabic. It employs an n-gram speech recognition engine, a rudimentary phrase-based translator for translating recognized Iraqi text, and a rudimentary text-to-speech (TTS) synthesis engine for playing back the English translation. This paper describes the system architecture, the functionality of its components, and the configurations of the speech recognition and machine translation engines.

1 Background

Humanitarian personnel, military personnel, and visitors in foreign countries often need to communicate with residents of a host country. Human interpreters are inevitably in short supply, and training personnel to speak a new language is difficult. Under the DARPA TRANSTAC and Babylon programs, various teams have developed systems that enable two-way communication over a language barrier (Waibel et al., 2003; Zhou et al., 2004; Stallard et al., 2006). The two-way speech-to-speech (S2S) translation systems seek, in principle, to translate any utterance, by using general statistical models trained on large amounts of speech and text data.

The performance and usability of such two-way speech-to-speech (S2S) translation systems is heavily dependent on the computational resources, such as processing power and memory, of the platform they are running on. To enable open-ended conversation these S2S systems employ powerful but highly memory- and computation-intensive statistical speech recognition and machine translation models. Thus, at the very minimum they require the processing and memory configuration of common-of-the-shelf (COTS) laptops.

Unfortunately, most laptops do not have a form factor that is suitable for mobile users. The size, weight, and shape of laptops render them unsuitable for handheld use. Moreover, simply carrying the laptop can be infeasible for users, such as military personnel, who are already overburdened with other equipment. Embedded platforms, on the other hand, offer a more suitable form factor in terms of size and weight, but lack the computational resources required to run more open-ended 2-way S2S systems.

In previous work, Prasad et al. (2007) reported on the development of a S2S system for Windows Mobile based handheld computers. To overcome the challenges posed by the limited resources of that platform, the PDA version of the S2S system was designed to be more constrained in terms of the ASR and MT vocabulary. As described in detail in (Prasad et al., 2007), the PDA based S2S system configured for English/Iraqi S2S translation delivers fairly accurate translation at faster than real-time.

In this paper, we present ongoing development work on an S2S system that runs on an even more constrained hardware platform; namely, a processor embedded in a wearable headset with just 32 MB of memory. Compared to the PDA based sys-

Proceedings of the ACL-08: HLT Workshop on Mobile Language Processing, pages 10–12,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

tem described in (Prasad et al., 2007), the wearable system is designed for both eyes- and hands-free operation. The headset-integrated translation device described in this paper is configured for two-way conversation in English/Iraqi. The target domain is the force protection, which includes scenarios of checkpoints, house searches, civil affairs, medical, etc.

In what follows, we discuss the hardware and software details of the headset-integrated translation device.

2 Hardware Platform

The wearable S2S system described in this paper runs on a headset-integrated computational platform developed by Integrated Wave Technologies, Inc. (IWT). The headset-integrated platform employs a 200 MHz StrongARM integer processor with a total of just 32MB RAM available for both the operating system and the translation software. The operating system currently running on the platform is Embedded Linux.

There are two audio cards on the headset platform for two-way communication through separate audio input and output channels. The default sound card uses the headset integrated close-talking microphone as an audio input and the second audio card can be used with an ambient microphone mounted on the device or an external microphone. In addition, each headset earpiece contains inner and outer set of speakers. The inner earpiece speakers are for the English speaking user who wears the headset, whereas the outer speakers are for the foreign language speaker who is not required to wear the headset.

3 Software Architecture

Depicted in Figure 1 is the software system architecture for the headset-integrated wearable S2S system. We are currently using a fixed-phrase English-to-Iraqi speech translation module from IWT for translating from English to Iraqi. In the Iraqi-to-English (I2E) direction, we use an n-gram ASR engine to recognize Iraqi speech, a custom, phrase-based "micro translator" for translating Iraqi text to English text, and finally a TTS module for converting the English text into speech. The rest of this paper focuses on the components of the Iraqi-to-English translation module.

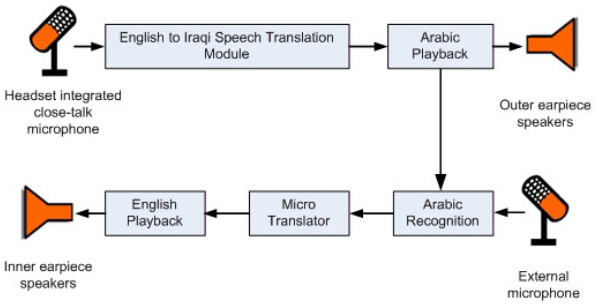

Figure 1. Software architecture of the S2S system.

Fixed point ASR Engine: The ASR engine uses phonetic hidden Markov models (HMM) with one or more forms of the following parameter tying: Phonetic-Tied Mixture (PTM), State-Tied Mixture (STM), and State-Clustered-Tied Mixture (SCTM) models.

For the headset-integrated platform, we use a fixed-point ASR engine described in (Prasad et al., 2007). As in (Prasad et al., 2007) for real-time performance we use the compact PTM models in both recognition passes of our two-pass ASR decoder.

Phrase-based Micro Translator: Phrase-based statistical machine translation (SMT) has been widely adopted as the translation engine in S2S systems. Such SMT engines require only a large corpus of bilingual sentence pairs to deliver robust performance on the domain of that corpus. However, phrase-based SMT engines require significant amount of memory, even when configured for medium vocabulary tasks. Given the limited memory on the headset platform, we chose to develop instead a phrase-based "micro translator" module, which acts like a bottom-up parser. The micro-translator uses translation rules derived from our phrase-based SMT engine. Rules are created automatically by running the SMT engine on a small training corpus and recording the phrase pairs it used in decoding it. These phrase pairs then become rules which are treated just as though they had been written by hand. The micro translator currently makes no use of probabilities. Instead, as shown in Figure 2, for any given Arabic utterance, the translator greedily chooses the longest matching source phrase that does not overlap a source phrase already chosen. The target phrases for these source phrases are then output as the translation. These target phrases come out in source-language

order, as no language model is currently used for reordering.

The micro translator currently consists of 1300 rules and 2000 words. Its memory footprint is just 32KB. This small memory footprint is achieved by representing the rules in binary format rather than text format.

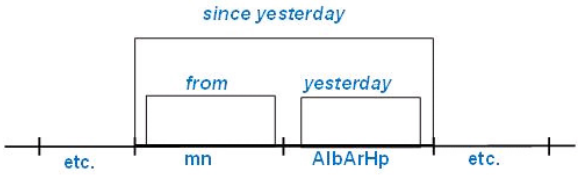

Figure 2. Decoding in micro translator.

English Playback using TTS: To play the English translation to the headset user we developed a rudimentary TTS module. The TTS module parses the output of the I2E translator to extract each translated word. It then uses the list of extracted words to read the appropriate pre-recorded (or synthesized) audio. Once the word pronunciations audio files are read we splice the beginning and the end of the audio files to reduce the amount of silence and concatenate them into a single file which is then played to the user on the inner earphone speakers.

The total memory footprint of our current Iraqi to English translation module running on the headset-integrated platform is just 9MB. The current configuration of the translation module's Iraqi ASR engine yields word error rate (WER) of 20% on test-set utterances without out-of-vocabulary (OOV) words.

4 Conclusions and Future Work

In this paper we have presented the initial setup of a speech-to-speech translation system configured for the headset platform. Our current work is focused on expanding the vocabulary of the Iraqi-to-English translation module by exploiting the rich morphology of Iraqi Arabic. In particular, we are investigating the use of morphemes (prefix, stems, and suffixes) for expanding the effective vocabulary of the headset translator. We are also developing use cases for performing a formal evaluation of both the usability and performance of the headset translator.

References

Alex Waibel, Ahmed Badran, Alan W Black, Robert Frederking, Donna Gates ,Alon Lavie, Lori Levin, Kevin Lenzo, Laura Mayfield Tomokiyo, J¨urgen Reichert, Tanja Schultz, Dorcas Wallace, Monika Woszczyna and Jing Zhang. 2003. "Speechalator: Two-way Speech-to-Speech Translation on a Consumer PDA," Proc. 8th European Conference on Speech Communication and Technology (EUROSPEECH 2003), Geneva, Switzerland.

Bowen Zhou, Daniel D´echelotte and Yuqing Gao. 2004. "Two-way Speech-to-Speech Translation on Handheld Devices," Proc. 8th International Conference on Spoken Language Processing, Jeju Island, Korea.

David Stallard, Frederick Choi, Kriste Krstovski, Prem Natarajan and Shirin Saleem. 2006. "A Hybrid Phrase-based/Statistical Speech Translation System," Proc. The 9th International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP), Pittsburg, PA.

David Stallard, John Makhoul, Frederick Choi, Ehry Macrostie, Premkumar Natarajan, Richard Schwartz and Bushra Zawaydeh. 2003. "Design and Evaluation of a Limited two-way Speech Translator," Proc. 8th European Conference on Speech Communication and Technology (EUROSPEECH 2003), Geneva, Switzerland.

Rohit Prasad, Kriste Krstovski, Frederick Choi, Shirin Saleem, Prem Natarajan, Michael Decerbo and David Stallard. 2007. "Real-Time Speech-to-Speech Translation for PDAs," Proc. IEEE International Conference on Portable Information Devices (IEEE Portable 2007), Orlando, FL.