Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 592–618

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

The IBM speech-to-speech translation system for smartphone: Improvements for resource-constrained tasks☆

Bowen Zhou*, Xiaodong Cui, Songfang Huang, Martin Cmejrek, Wei Zhang, Jian Xue, Jia Cui, Bing Xiang, Gregg Daggett, Upendra Chaudhari, Sameer Maskey, Etienne Marcheret

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, United States

Received 15 February 2010; received in revised form 22 August 2011; accepted 30 August 2011
Available online 21 September 2011

Abstract

This paper describes our recent improvements to IBM TRANSTAC speech-to-speech translation systems that address various issues arising from dealing with resource-constrained tasks, which include both limited amounts of linguistic resources and training data, as well as limited computational power on mobile platforms such as smartphones. We show how the proposed algorithms and methodologies can improve the performance of automatic speech recognition, statistical machine translation, and text-to-speech synthesis, while achieving low-latency two-way speech-to-speech translation on mobiles.

Keywords: Speech-to-speech translation; Speech recognition; Machine translation; Text-to-speech; Low-resourced languages; Mobiles

1. Introduction

In today's global world, there are obvious ever-increasing needs for bridging the barriers between major languages (TC-STAR; Sakti et al., 2009) for various social and economic activities. In addition, there is often a surging demand for speech-to-speech (S2S) technologies that address some less prominent languages, partially due to the lack of human translators. And more often, pressing needs for automatic S2S translation arise from abrupt events such as the international relief efforts following the 2010 Haiti earthquake. Under such scenarios, users typically have a strong need to access S2S technologies anytime and anywhere, ideally on mobiles, such as smartphones, that may or may not have connections to the Internet.

The challenges of developing a two-way S2S system for such tasks are twofold. Firstly, there are usually a very limited amount of required training data (such as audio recordings with transcriptions and parallel text data between source and target languages) and linguistic resources available; In addition, the research team often lacks language-specific linguistic expertises in-house, especially at the beginning of the development. Secondly, resources are computationally constrained on mobiles with a limited amount of run-time memory and CPU power, while in this task, a complex

☆ This paper has been recommended for acceptance by Guest Editors Speech-Speech Translation.

* Corresponding author. Tel.: +1 914 945 2852.
E-mail address: zhou@us.ibm.com (B. Zhou).

0885-2308/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2011.08.004

two-way S2S system typically including multiple computationally intensive component modules, is expected to be completely hosted and run smoothly with low-latency on such mobile platforms.

The Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) program (DARPA; Gao et al., 2006; Stallard et al., 2007; Precoda et al., 2007; Bach et al., 2009; Belvin et al., 2005) aims to address these challenges. Built upon prior efforts to develop wearable Iraqi-English S2S systems (e.g., with the hardware platforms built around the Panasonic U1) under TRANSTAC, the program was pushing the limits by taking on the following challenges: (1) to develop self-contained end-to-end two-way S2S systems for smartphone devices (e.g., iPhone or Nexus One), and (2), to rapidly develop the S2S capability for Dari and Pashto, two languages with limited resources.

In this paper, we will first give an overview of the key technologies that we have employed to develop the wearable S2S systems. This serves as the starting point to continue our efforts for the smartphone-based system. We will also review how the S2S system was evaluated in the TRANSTAC evaluation. Next, in the core of this paper, we will present our recent improvements to IBM TRANSTAC speech-to-speech translation systems that address various issues arising from dealing with resource-constrained tasks, which include both limited amounts of linguistic resources and training data, as well as limited computational power on smartphones. We will show how the proposed algorithms and methodologies can improve the performance of key components such as automatic speech recognition (ASR), statistical machine translation (SMT) and text-to-speech synthesis (TTS), while achieving low-latency two-way speech-to-speech translation on smartphones.

Although some of the methods and technologies discussed in this paper can be applied to multiple components, the paper is presented by organizing the topics around the three key areas of ASR, SMT and TTS. We have tried to make each topic as self-explanatory as possible by directly enclosing experimental results and pointers to references and further details when applicable.

The rest of the paper is structured as follows: Section 2 outlines the IBM TRANSTAC system as a wearable S2S solution, which serves as the baseline system for our continued discussion; Section 3 presents various ASR strategies for improved acoustic modeling of sparse data, as well as the transformation compression and streaming architecture developed for the runtime decoder on smartphones to achieve low-latency, low-resource recognition; Section 4 introduces several methods that can effectively improve the translation performance for synchronous context-free grammar based SMT systems; Section 5 gives an overview of our TTS module and shows how we rapidly build high-quality voices for low-resourced languages; in particular, we will describe our efforts to leverage several machine learning techniques used in our ASR and SMT systems to reduce our dependency on language specific resources; and finally, Section 6 will summarize our contributions in this paper.

2. The IBM wearable speech-to-speech translation system for TRANSTAC

This section reviews the IBM TRANSTAC S2S system for Iraqi Arabic-English, which is made wearable for users by running on a ruggedized tablet computer. The technologies described below serve as the baseline on which we strive to improve for more sparse languages such as Dari and Pashto (details presented in Sections 3–5) on smartphone platforms.

We will present the official IBM results from the TRANSTAC June and November 2008 evaluations, which shed light on how the end-to-end system performs from a task-oriented perspective. We will also include objective measurements of component performance and show that, as expected, the improvements in a component often lead to better overall end-to-end performance.

2.1. Overview

We will describe the overall structure of the ASR and SMT components, leaving the TTS component to be presented in Section 5, where the performance of different voices developed using a similar technique is compared.

2.1.1. ASR

The construction of the ASR component of the IBM S2S translation system is illustrated in Fig. 1, including feature extraction, discriminative training in both feature and model spaces, and streaming decoding on a static graph.