Multilingual Web Conferencing Using Speech-to-Speech Translation

John Chen, Shufei Wen, Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore

AT&T Labs - Research
180 Park Avenue, Florham Park, NJ 07932, USA
jchen,shufei,vkumar,srini@research.att.com

Abstract

It is now commonplace to use web conferencing technology in order to hold meetings between participants situated in different physical locations. A drawback of this technology is that nearly all of the interaction between these participants is monolingual. Here, we demonstrate a novel form of this technology that enables cross-lingual speech-to-speech communication between conference participants in real time. We model this translation problem as a combination of incremental speech recognition and segmentation, addressing the question of finding which segmentation strategy maximizes translation accuracy while minimizing latency. Our demonstration takes the form of a web conferencing scenario where a presenter speaks in one language while talk participants listen to or read the speaker's translated texts in real time. This system is flexible enough to allow real-time translation of technical talks or speeches covering broad topics.

Index Terms: Web conferencing, simultaneous speech-to-speech translation, session initiation protocol (SIP)

1. Introduction

Web conferencing technology has advanced significantly that it is now commonly used to connect users (employees or general public) in remote locations. The services provided by such technology include online chat applications (Microsoft Skype, Google Hangout); meetings, webinars and remote technical support (Citrix GoToAssist, GotoMeeting, Cisco WebEx, etc.). Most of these services contain a speech, text and video channel for enabling communication across the participants. However, almost all the interaction between the participants is monolingual; i.e., audio or text is merely transmitted from one place to another. Enabling cross-lingual communication using speech-to-speech translation technology can potentially connect participants across languages and also provide an automatic way of archiving/searching meetings using indexed speech translation output.

The nature of interaction through web conferencing demands that the communication be seamless and real-time. As a result, multilingual web conferencing using speech-to-speech (S2S) or speech-to-text (S2T) technology needs to be real-time with low latency and high accuracy. Furthermore, the participants do not take well defined explicit turns and can talk continuously for long periods of time. Simultaneous S2T translation has been addressed before in the TC-STAR project for translating European Parliamentary speeches as well as lectures [1, 2, 3, 4]. However, the work focused on a single speaker (with offline adaptation) in a limited domain. Simultaneous S2S translation was presented in [5] for a two-way interaction with explicit turn taking, and more recently for translation of TED talks [6] where speaker(s) can talk continuously for long stretches of time. We repurpose the speech translation models trained in [6] for the web conferencing setting described in this work.

We model the simultaneous S2S translation problem as a combination of incremental speech recognition and segmentation; i.e., we address the question, "what is the segmentation strategy that maximizes the number of segments while still maximizing translation accuracy?". Our experimental results in [6] indicated that performing speech translation on sub-sentential comma-separated chunks results in the highest accuracy with lowest latency. Hence, we used the same strategy in the web conferencing setting. Our demonstration addresses two different scenarios for a web conferencing application: (i) Spanish speaker speaking and participants receiving English text and audio in real-time, (ii) English speaker speaking and participants receiving real-time Spanish text and audio. We provide the ability to choose models for either technical talks (internal talks) or speeches covering broad topics (TED talks). The entire communication is facilitated using a session initiation protocol (SIP).

In the following section, we briefly describe the system as well the components used in our demonstration. A more comprehensive description of the models is presented in [6].

2. System Description

In order to implement real-time speech translation with text message display, a Back-To-Back User Agent (B2BUA) application server (app server) was used to act as a SIP Registrar and Soft-Switch was used to connect calls to AT&T WATSON^{SM} SIP servers. The WATSON SIP servers perform the S2S translation, i.e., they use the specified models and run through the pipeline of speech recognition, segmentation, machine translation and text-to-speech synthesis. The B2BUA app server was implemented using AT&T Echarts for SIP Servlets (E4SS), which is a software for finite state machine programming that provides an easy to use B2BUA framework. The SIP user agent (UA) for the speaker dials the SIP app server with the number "9010" for English to Spanish or the number "9011" for Spanish to English. The app server finds the URL for the WATSON SIP server that corresponds to the number dialed using its registration database. It then sends another SIP:INVITE message to that WATSON with session description protocol (SDP) content that is modified from the original SIP:INVITE message. After all the SIP:INVITE negotiations are completed, the audio-encoded real-time transport protocol (RTP) packets flow between WATSONs and the users' end devices directly. For each translation, the WATSON server sends the translated text using SIP:MESSAGE to the app server which in turn forwards it to the user's end device. An illustration of the setup and call flow is shown in Figure 1.

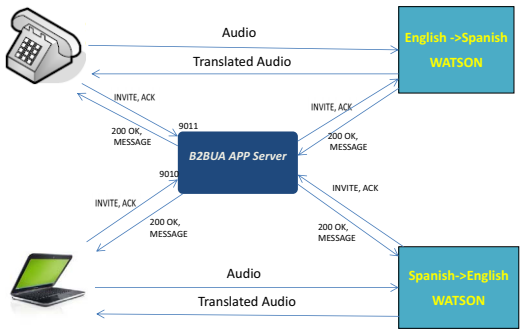

Figure 1: Diagram of system that performs real-time speech translation accompanied by text message displays.

2.1. Speech Recognition

We use the AT&T WATSON^{SM} speech recognizer [7]. The English acoustic model is trained on just over 600 hours of speech obtained from TED Talks [1]. The Spanish acoustic model is trained on over 1000 hours of speech collected through smartphones. The baseline model training was initialized with several iterations of ML training, including two builds of context dependency trees, followed by three iterations of Minimum Phone Error (MPE) training. Our framework currently does not perform any online adaptation.

The English language model was built using the permissible data in the IWSLT 2011 evaluation. The texts were normalized using a variety of cleanup, number and spelling normalization techniques and filtered by restricting the vocabulary to the top 375000 types. We used the AT&T FSM toolkit [8] to train a trigram language model (LM) for each component (corpus). Finally, the component language models were interpolated by minimizing the perplexity on a development set. The Spanish language model was built in a similar way over the spanish side of the parallel corpora.

2.2. Segmentation

The segmenter predicts the locations of commas and ends of sentences in a stream of words that is input from the speech recognizer. It employs a logistic regression model [9] which is trained on a subset of transcribed and translated TED talks. After segmentation, each word segment, separated by commas or periods from other segments, is translated individually. In addition, the output of the segmenter is used to pretty print the input English text as well as its corresponding Spanish translation.

The training data contains 1029 human-transcribed and translated talks. This parallel data consists of about 150,000 sentences in both English and Spanish. Punctuation in this data was normalized so that only commas and sentence-ending periods appeared. Specifically, different kinds of sentence ending punctuations were transformed into a uniform end of sentence marker. Double-hyphens were transformed into commas. Commas already existing in the input were kept while all other kinds of punctuation symbols were deleted.

[1] http://www.ted.com

Part of speech (POS) taggers were applied to the data, one for English and another for Spanish. For speed, unigram POS taggers were implemented. They used word shape features to predict the POS of unknown words. The English POS tagger was trained on Sections 02-21 of the Penn Treebank [10], about 950,000 words. The Spanish POS tagger was trained on the output of the Freeling POS tagger [11] on the Spanish part of the English-Spanish portion of the European Parliamentary Interpretation Corpus (EPIC) [12], consisting of about 50,000 words.

The segmenter relies on a five word and POS window in order to classify the target word. Specifically, token t_0 is classified given as input the window $t_{-2}t_{-1}t_0t_1t_2$. Unigram, bigram, and trigram word and POS features based on this window were used for classification.

2.3. Machine Translation

We used the Moses toolkit [13] for training a phrase-based statistical machine translation model. Subsequently, we represent the phrase translation table as a weighted finite state transducer (FST) and the language model as a finite state acceptor (FSA). The weight on the arcs of the FST is the dot product of the MERT weights with the translation scores. In addition, a word insertion penalty was also applied to each word to penalize short hypotheses. The decoding process consists of composing all possible segmentations of an input sentence with the phrase table FST and language model, followed by searching for the best path. Our FST-based translation is the equivalent of phrase-based translation in Moses without reordering. The parallel data used for building the models is described in [6].

2.4. Text-to-Speech Synthesis

The translated sentence from the machine translation component is synthesized using the AT&T Natural Voices^{TM} text-to-speech synthesis engine [14]. The system uses unit selection synthesis with half phones as the basic units. It is sensitive to punctuation marks predicted by the segmenter, inserting short pauses whenever a comma was predicted and longer pauses whenever an end of sentence was predicated. The database was recorded by professional speakers of the language. We use either male or female voice for English and Spanish depending on the source language.

3. Conclusion

We demonstrate a web conferencing application having the enhanced capability of simultaneous S2S translation. Two use case scenarios include (i) Spanish speaker speaking and participants receiving English text and audio in real-time, (ii) English speaker speaking and participants receiving real-time Spanish text and audio. In this application, an app server is responsible for setting up SIP channels connecting speaker, participants, and WATSON servers performing S2S translation. Once these channels are established, S2S translation is performed using a pipeline consisting of AT&T WATSON speech recognizer, Text Segmenter, phrase-based FST MT, and TTS through AT&T Natural Voices. This system is flexible enough to allow real-time translation of technical talks or speeches covering broad topics.

4. References

[1] D. Vilar, E. Matusov, S. Hasan, R. Zens, and H. Ney, "Statistical machine translation of European parliamentary speeches," in *Proceedings of MT Summit*, 2005.

[2] C. Fügen, M. Kolss, D. Bernreuther, M. Paulik, S. Stuker, S. Vogel, and A. Waibel, "Open domain speech recognition & translation: Lectures and speeches," in *Proceedings of ICASSP*, 2006.

[3] O. Hamon, C. Fügen, D. Mostefa, V. Arranz, M. Kolss, A. Waibel, and K. Choukri, "End-to-end evaluation in simultaneous translation," in *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, March 2009.

[4] C. Fügen, A. Waibel, and M. Kolss, "Simultaneous translation of lectures and speeches," *Machine Translation*, vol. 21, pp. 209–252, 2007.

[5] S. Bangalore, V. K. Rangarajan Sridhar, P. Kolan, L. Golipour, and A. Jimenez, "Real-time incremental speech-to-speech translation of dialogs," in *Proceedings of NAACL:HLT*, June 2012.

[6] V. K. Rangarajan Sridhar, J. Chen, S. Bangalore, A. Ljolje, and R. Chengalvarayan, "Segmentation strategies for streaming speech translation," in *Proceedings of NAACL:HLT*, 2013, To appear.

[7] V. Goffin, C. Allauzen, E. Bocchieri, D. Hakkani-Tür, A. Ljolje, and S. Parthasarathy, "The AT&T Watson Speech Recognizer," Tech. Rep., September 2004.

[8] M. Mohri, F. Pereira, and M. Riley, "At&t general-purpose finite-state machine software tools, http://www.research.att.com/sw/tools/fsm/," 1997.

[9] P. Haffner, G. Tür, and J. Wright, "Optimizing svms for complex call classification," in *Proceedings of ICASSP'03*, 2003.

[10] M. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of English: the Penn treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[11] L. Padró and E. Stanilovsky, "Freeling 3.0: Towards wider multilinguality," in *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. Istanbul, Turkey: ELRA, May 2012.

[12] C. Bendazzoli and A. Sandrelli, "An approach to corpus-based interpreting studies," in *Proceedings of the Marie Curie Euroconferences MuTra: Challenges of Multidimensional Translation*, Saarbrücken, 2005.

[13] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, S. W., C. Moran, R. Zens, C. J. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of ACL*, 2007.

[14] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS system," in *Proceedings of Joint Meeting of ASA, EAA and DEGA*, 1999.