Speech to speech translation: a communication boon

Karunesh Arora · Sunita Arora · Mukund Kumar Roy

Received: 1 November 2012 / Accepted: 16 April 2013 / Published online: 7 June 2013
© CSI Publications 2013

Abstract An average person speaks 11000–25000 words per day making speech the most common way of expressing ourselves. Be it a conversation, dialogue, speech, presentations or any general talks, we use speech to make other as well as ourselves understand thoughts and actions. If either of the side is unaware of the language of communication, the cycle will be incomplete. Hence we need a system that can bridge this language barrier. Speech to speech translation is one such system that can play important role by facilitating communication between persons speaking different languages. Worldwide efforts are being made to achieve this goal and implement it practically for use by common man. The present paper describes a major international and inter-institutional effort in this direction—in which an attempt is being made to automate speech translation among 23 Asian, Middle East and European languages including Hindi through a consortium project led by NICT Japan [1, 2]. The three key modules namely Speech Recognition, Language Translation and Speech Synthesis required for Hindi are being designed, developed and implemented by CDAC, Noida as Indian counterpart in the project. The language specific technology and the parallel corpora and the speech unit (segmental) database developed have been described. Technical details of this first ever effort, modules and their

performance in the communication system have been discussed.

Keywords Speech to speech translation · ASR · Statistical MT · TTS · U-STAR

1 Introduction

Owing to this era of global and border less economy, the information exchange has become inevitable and speech is the traditional and best way for the same. The global scenario adds the demand of communication among speakers of different languages. This is actually one of the great challenges before the Information Technologists to overcome language barriers across the global community, and enable them to express themselves in real time. Further, the task of bridging the Digital-divide can never be accomplished in real sense without breaking the language barriers with an intelligent system or machines. Speech translation technology—being able to speak and have one's words translated automatically into the other person's language—has long been a dream of humankind. In the recent predictions, speech to speech translation has been placed as one of the ten technologies that will change the world.

All the languages prevailing in our world have different origins and set of their native mother tongue is also quite discrete. According to a study, it is quite cumbersome and costly for an 'adult' to learn new language than a child. Almost all worldly population finds acquisition of foreign languages extremely difficult due to the geographical factors. Thus speech to speech translation technology would be a great boon to all.

Manual translation has been limited to important official documents, news items and some award winning literary

K. Arora (✉) · S. Arora · M. K. Roy
NLP Lab, CDAC, Anusandhan Bhawan, C-56/1, Sector 62,
Institutional Area, Noida 201307, India
e-mail: karunesharora@cdac.in

S. Arora
e-mail: sunita@cdac.in

M. K. Roy
e-mail: mukundkumarroy@cdac.in

Table 1 Major efforts in S2S Translation

	Research	Commercial	Military
US	Janus, DIPLOMAT, Tongues, Nespole! Maxtor	AT & T	Phraselator, Babylon LASER
Europe	Verbmobil, Nespole! LC-Star, TC-Star		
Japan	MATRIX	NEC	
China	LodeStar, Digital Olympics		

works. There exists a huge backlog of materials that needs to be translated for administration, education, commerce, tourism etc. Technological support in the form of machine aids for translation is of great importance.

2 Background

Speech-translation technology is significant because it enables speakers of different languages from around the world to communicate, erasing the language divide in global business and cross-cultural exchange. Achieving speech translation would have tremendous scientific, cultural, and economic value for humankind. The article “10 Emerging Technologies That Will Change Your World” in the issue of An MIT Enterprise Technology Review lists “Universal Translation” as one of these ten technologies.

In 2007 CDAC entered into an agreement for participating in the consortium led by ATR, NICT Japan titled as Asian Speech Translation Advanced Research (A-STAR) along with research institutions from China, Korea, Thailand, Indonesia and Vietnam. In July 2009, first Asian Network based S2S system was launched. Later this consortium expanded with introduction of Middle East and European countries. Now this initiative is known as Universal Speech Translation Advance Research (U-STAR). The objective of this consortium is to cooperate in research in automated speech translation in Asian, Middle East and European languages by conducting joint research.

3 History and world scenario of S2S translation systems

Speech translation was first noticed during 1983 ITU Telecom World (Telecom’83), when NEC Corporation made a demonstration of speech translation as a proof of concept. In 1993, an experiment in speech translation was conducted linking three sites around the world: the ATR, Carnegie Mellon University (CMU) and Siemens. Germany launched the Verbmobil project; the European Union the Nespole! and TC-Star projects; and the United States launched the TransTac and GALE projects [3].

Research and development has gradually progressed from relatively simple to more advanced translation, progressing

from scheduling meetings, to hotel reservations, to travel conversation. Moving forward, however, there is a need to further expand the supported fields to include a wide range of everyday conversation and sophisticated business conversation.

There have been several efforts of developing S2S Translation systems and there have been some success stories. The Table 1 given in the annexure lists the major efforts in S2S Translation.

4 Architecture for multilingual speech to speech translation

The Universal Speech Translation Advanced Research Consortium (U-STAR) is an international research collaboration entity formed to develop a network-based speech-to-speech translation (S2ST) with the aim of breaking language barriers around the world and to implement vocal communication between different languages. In 2010, as an initiative of U-STAR, international communication protocols were standardized and approved by the ITU-T; Recommendations F.745 [4], and H.625 [5], enabling speech-to-speech translation (S2ST) modules to be connected across the globe over networks. Below are the Excerpts from these recommendations:

4.1 Network-based speech-to-speech translation services

This Recommendation specifies the service description and the requirements for speech-to-speech translation (S2ST) accomplished by connecting distributed S2ST modules all over the world through a network. This service provides S2ST that recognizes the speech in one language, translates the recognized speech into another language, and then synthesizes the translation into speech. People who speak different languages can communicate using this service.

The applications and services using network-based S2ST technologies are characterized by the following components:

4.1.1 S2ST client

- User client for speech/text input and output.

Fig. 1 Network-based S2S translation system

4.1.2 S2ST servers

- Speech recognition: speech is recognized and transcribed;
- Machine translation: text in source language is translated into text in target language;
- Speech synthesis: speech signal is created from text.

4.1.3 Communication protocol

- Communication protocol to connect user clients and the above S2ST servers.

In order to extend the network-based S2ST to other modalities (e.g., sign language), a communication protocol is incorporated for modality conversion (MC), which converts single/multiple modality information to different single/multiple modality information. The communication protocol for MC needs to have an expandable structure.

4.1.4 Modality conversion markup language (MCML)

- XML schema that serves as a data description for data exchanged among modality conversion modules.

The leveraging of S2ST technologies in a pragmatic manner, which has long been one of mankind's dreams, may have a significant impact on tourism, social services, safety, and security. To construct S2ST systems, automatic speech recognition (ASR), machine translation (MT) and text-to-speech synthesis (TTS) must be built for source and target languages by collecting speech and language data, such as

audio data, its manual transcriptions, pronunciation lexica for each word, parallel corpora for translation and so on.

It is very difficult for individual organizations to build S2ST systems covering all topics and languages. However, by interconnecting ASR, MT and TTS modules developed by separate organizations and distributed globally through a network, one can create S2ST systems that break the world's language barriers.

The consortium is working collaboratively to collect language corpora, create common speech recognition and translation dictionaries, develop Web service speech translation modules for the various Asian languages, and standardize interfaces and data formats that facilitate the international interaction among the different speech translation modules from different countries.

The system is being designed to translate common spoken utterances of travel conversations from a certain source language into multiple target languages in order to facilitate multiparty travel conversations between people speaking different languages. Each consortium member contributes one or more of the following spoken language technologies: automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS) through Web servers. Currently, the system is covering 23 languages aiming 23*22/2 S2S systems.

Figure 1 illustrates an example where a spoken Source Language utterance is recognized and converted into Source Language text; this is then translated into Target Language text, which is synthesized into Target Language speech [1].

In a loosely coupled system, these three components are connected in a sequential style: ASR converts user's speech to text in source language and then MT translates

Fig. 2 Protocols of network-based speech-to-speech translation

the source text into the target language. Finally TTS creates the synthesized speech from the target text.

5 System architecture

5.1 Client application

This application is a multi-device application available for Free for iPhone as well as Android phones “VoiceTra4U-M” [6]. It helps multiple users (up to 5) communicate in different languages, in real time either face to face or remotely.

The application contributes to breaking the barriers of modalities other than language as well. For instance, it helps users to communicate with the visually-impaired via spoken word, or with the hearing-impaired via text input.

5.2 Control servers

Control server is used to relay the speech results from one user to all other users in order to enable them to perform a multiparty conversation as well as dialogue based Conversation. Control Server at the target side decides which service (ASR, MT or TTS) to invoke and send response back.

5.3 Communication servers

Communication Servers are designed, implemented and maintained by their respective language verticals. These

servers actually perform ASR, MT and TTS service based on the “service invoke request” by the Control Server.

Figure 2 below illustrates Protocols of Network-based Speech-to-Speech Translation and Fig. 3 is the glimpse of real time scenario of S2S Translation.

6 Technology

6.1 Automatic speech recognition (Hindi)

A technical definition of ASR is the building of system for mapping acoustic signals to a string of words. The general problem of automatic transcription of speech by any speaker in any environment is still far from solved. But recent years have seen ASR technology mature to the point where it is viable in certain limited domains.

Five main sub-components of an ASR system are [7–9]

- Acoustic Model (p(O|W))
- Language Model (P(W))
- Lexicon/Pronunciation Model (HMM)
- Feature Extraction
- Decoder

For building Acoustic model, we used audio data of 8567 sentences that accounted for more than 60 h of recording. These sentences were recorded in a clean noise free environment, by the speakers uniformly distributed over all age groups from 17 to 60 years. Prototype models

Fig. 3 Working of S2S translation system

for 61 phonemes are built using flat start approach. The hmm models are over 61 context independent phonemes. Julius recognition engine is used for decoding the utterances which is a two pass stack decoder. The performance of LVCSR is measured in terms of recognition rate. The system is tested on

1. 10 seen speakers
2. 10 unseen speakers

Table 2 shows performance of Hindi ASR on both Seen and Unseen Speakers [10].

6.2 Machine translation (English–Hindi, Hindi–English)

The approach for implementation is primarily using statistical machine translation (SMT). The advantage of SMT is that one does not require deeper syntactic understanding of Source and Target languages. The base models can be quickly built, as soon as we have the parallel corpus of the language pair with us. In the proposed scenario it is very difficult to have the man-power who is having experience and knowledge of multiple languages. And the languages of the consortia members are not even remotely related. Hence, SMT was the obvious choice for development. English has been chosen as the linking language around which each consortia member is developing their own language corpus. Then the respective translation models are to be developed among different language pairs directly without intervening English in between them. Each consortia member is tuning the system by supplementing linguistic information, such as transliterations, part-of-speech

and chunk information. The Fig. 4 below shows the basic architecture of the typical SMT system.

MT for En–Hi and Hi–En pairs were trained on multi-lingual BTEC (Basic Travel Expressions Corpus) Corpus which is a set of 20 K sentences. The challenge in these language pairs lies in their difference in word order—Hindi being SOV (subject–object–verb), while English observing SVO (subject–verb–object) and degree of inflections. Hindi is morphologically richer and more inflectional in nature than English.

The Corpus was then subdivided into three sets namely Training (Trg), Development (Dev) and Test (Tst) sets containing 18000, 1000, 1000 sentences respectively. For

Fig. 4 Statistical machine translation**Table 2** Word Recognition rate

Speaker	Sentences	Recognition rate
Training	3128	70.73 %
Test	3128	60.66 %

Table 3 Different statistics observed

# Sentences		Trg	Dev	Tst	BLEU score	
					Eng	Hin
En	# words	153066	17806	18257	–	0.1596
	#Avg. words	7.7	7.6	7.7		
Hi	# words	171347	17774	17811	0.1785	–
	#Avg. words	8.6	8.7	8.6		

Fig. 5 Concatenative speech synthesis

the training of MT model, Open source learning algorithms were used.

Decoding parameters were tuned using MERT and finally translations were performed using phrase based decoder. Table 3 gives insight of different statistics observed while building SMT [11].

6.3 Text to speech synthesis (Hindi)

The Festival speech synthesis system is primarily designed for phoneme and di-phone units, and we adapted it to work for syllable units. For data preparation, the text prompts were recorded in an anechoic chamber by a professional speaker. The recorded prompts were manually labelled at phoneme level using EMU speech tool. The prompts were further labelled to higher levels like syllables and words. The prosodic phrasings were also introduced in the database. The text processing module breaks the incoming text sentence to a syllable sequence. The unit selection module selects the best unit realization sequence from the many possible unit realization sequences for the given syllable sequence. The prosody prediction module predicts energy, pitch etc. Finally, in the concatenation module, the units are modified according to the predicted prosody before concatenation. Figure 5 illustrates a functional TTS.

7 Standardization

Other than the research and development of technology, the consortium's objective is to establish an international joint-research organization to design formats of bilingual corpora that are essential to advance the research and development of this technology, to design and compile basic bilingual corpora between languages, and to standardize interfaces and data formats to connect speech translation modules over the Internet jointly with research institutions working in this field.

It is also necessary to create common speech-recognition and translation dictionaries, and compile standardized bilingual corpora. The basic communication interface will be Web-based that will comply with international communication protocols standardized and approved by the ITU-T so that S2S modules are connected across the globe over networks.

8 Conclusion

The first major technology demonstration showcasing feasibility of the project was done in 2009 with the launch of first Network based S2S Translation system on a hand held device. In the next phase more consortium members including Middle Eastern and European languages. Standards for Network-based Communication were drafted. Consortium successfully launched Mobile application "VoiceTra4u-M" [6] in London prior to Olympic Games held there. Future tasks include improving upon and optimizing current technology. Collaborating with more and more countries and their language to form a global consortium, share research activities among the multilingual communities and making it accessible and usable for the target audience are the next milestones for this consortium.

Acknowledgments We would like to thank Dr. Eiichiro Sumita, Executive Director, NICT Japan for successfully leading the project and providing us the opportunity for being the part of this initiative. We would also thank Dr B K Murthy, ED, CDAC Noida for constant support and conducive environment for this work. And finally sincere thanks to all USTAR Consortia members.

References

1. Sakti S, Kimura N, Paul M, Hori C, Sumita E, Nakamura S, Park J, Wutiwathchai C, Xu B, Riza H, Arora K, Luong CM, Li H (2009) The Asian Network-based Speech-to-Speech Translation System, ASRU, 2009 and IEEE Explore, Nov. 13, 2009–Dec. 17, 2009, pp 507–512

2. Nakamura S, Sumita E, Shimizu T, Sakti S, Sakai S, Zhang J, Finch A, Kimura N, Ashikari Y (2007) A-STAR: Asia speech translation consortium. In: Proceedings ASJ Autumn Meeting, Yamanashi, Japan, 2007, pp. 45–46
3. Nakamura S (2009) Overcoming the language barrier with speech translation technology. Science and Technology Trends—Quarterly Review No. 31, Apr 2009, pp. 35–48
4. ITU-T; Recommendations. <http://www.itu.int/rec/T-REC-F.745-201010-I>
5. ITU-T; Recommendations. <http://www.itu.int/rec/T-REC-H.625-201010-I>
6. <http://itunes.apple.com/us/app/voicetra4u-m/id537952299?l=en&ls=1&mt=8>. Jan 2013
7. Young S, Ever G, Hain T, Kershaw D, Moore G, Odell J, Ollason D, Povey D, Vaitchev V, Woodland P “The HTK Book”, copyright 2001–2002 Cambridge University, Engineering Department
8. Lee A, Kawahara T, Shikano K Julius—an open source real-time large vocabulary recognition engine In: Proc. Eurospeech, pp. 1691–1694
9. Kumar M et al (2004) A large-vocabulary continuous speech recognition system for Hindi. IBM Res Dev J 48:703–715
10. Arora S, Saxena B, Arora KK, Agarwal SS (2010) Hindi ASR for Tourism domain In: Proc. O-COCOSDA 2010
11. Arora KK, Sinha RMK (2012) Improving statistical machine translation through co-joining verbal construct in English–Hindi machine translation. In: Proc. SSST-6, ACL, pp 95–101