A-STAR: Asian Speech-to-Speech Translation Research Consortium
- Towards Connecting Speech-Translation Systems in the Asian Region -

Satoshi Nakamura[1], Jun Park[2], Chai Wutiwiwatchai[3], Hammam Riza[4]
Bo Xu[5], Karunesh Arora[6], Chi Mai Luong[7], Haizhou Li[8]
[1]National Institute of Information and Communications Technology (NICT), Japan*
[2]Electronics and Telecommunications Research Institute (ETRI), Korea
[3]National Electronics and Computer Technology Center (NECTEC), Thailand
[4]Agency for the Assessment and Application of Technology (BPPT), Indonesia
[5]Institute of Automation, Chinese Academy of Sciences, China
[6]Center for Development of Advance Computing (CDAC), India
[7]Institute of Information Technology (IOIT), Vietnam
[8]Institute for Infocomm Research (I^{2}R), Singapore

Abstract

The new global, borderless economy has made it critically important for speakers of different languages to be able to communicate. Speech translation technology—being able to speak and have one's words translated automatically into the language of the person one is addressing—has long been a dream of humankind. Speech translation is regarded as one of the ten technologies that will change the world. The A-STAR (Asian Speech Translation Advanced Research) consortium was established in June 2006 by NICT/ATR (Japan) with the aim of realizing the objective of establishing network-based speech-to-speech translation systems in the Asian region. The A-STAR consortium originally comprised six countries: Japan, China, Korea, Thailand, Indonesia, and India. In 2008, Singapore and Vietnam also joined the A-STAR consortium. The consortium is presently working collaboratively toward collecting Asian language corpora, developing web-service modules for Asian languages, and advancing its common connection protocols and data formats. Small scale experiments of the network-based speech-to-speech translation system between the Japanese and Korean languages were conducted in March 2009. The A-STAR consortium is planning on conducting large scale network-based speech translation experiments in July 2009.

1 Introduction

Speech translation is a technology that translates one spoken language into another. Speech-translation technology is significant in today's world because it enables communication among speakers of different languages world-wide, thus erasing the inherent language divide in global businesses and cross-cultural exchanges. The achievement of speech translation holds tremendous scientific, cultural, and economic value for humankind. The article "10 Emerging Technologies That Will Change Your World" that appeared in the February 2004 issue of "An MIT Enterprise Technology Review" includes "Universal Translation" in the list of these ten technologies. While the article showcases a number of translation technologies, its focus is on speech-translation technology.

In recognition of the fact that many years of basic research would be required before the successful implementation of speech translation, the Advanced Telecommunications Research Institute International (ATR) was founded in 1986, where a project to research speech translation was begun. Researchers from a wide range of research institutes, located both in Japan and internationally, joined this project (Nakamura et al., 2006). The year 1991 saw the establishment of the C-STAR[1] consortium that aimed to coordinate various international speech-translation research activities through the ATR, the Carnegie Melon University (CMU), and Siemens. In 1993, a speech-translation experiment was conducted, linking three sites of the C-STAR consortium around the world. After the start of the ATR project, speech-translation projects were ini-

[*] The Spoken Language Communication Research Group of NICT was previously a part of ATR Spoken Language Communication Research Laboratories, Japan

[1] http://www.c-star.org

tiated around the world. Germany launched the Verbmobil[2] project (1993–1996); the European Union started the Nespole![3] (2000–2002) and TC-STAR[4] projects (2004–2007); and the United States launched the TransTac:[5] (2008) and GALE[6] projects (2006–). The GALE project was started in 2006 with the aim of automatically translating Arabic and Chinese into English. The goal of this ongoing project is to automate the extraction of vital multilingual information that has hitherto been performed exclusively by humans; the project architecture consists of a batch text-output system that has been established toward this purpose. In contrast, the objective of the ATR project is speech translation enabling face-to-face and non-face-to-face cross-language communication in real time. Online speech-to-speech translation is thus an integral component of this research, and immediacy of processing, a key factor.

Speech translation integrates three components: automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS). Each of these technologies presents its own difficulties. A particular requirement of this technology is the recognition and translation of spoken language; this objective is much more difficult to achieve than the translation of text because spoken language contains ungrammatical and colloquial expressions, in addition to not including punctuation like question marks, exclamation marks, or quotation marks. Mistakes committed in speech recognition can also cause major translation errors.

2 International collaboration for more languages in Asia

As speech-translation technology overcomes linguistic barriers, it would be ideal if researchers and research institutions from around the world could conduct collaborative research work on the subject. The C-STAR international consortium for joint research of speech translation has been quite active in bringing about such combined international research work.

Meanwhile, the travel destinations of international travelers—whether for the purpose of tourism, emigration, or foreign study–are becoming increasingly diverse. These and other changes are heightening the need for devising means that effect interaction among speakers of English and people from non-English-speaking countries.

In particular, the Asian region—including Russia—has witnessed a strengthening of its social and economic relationships. The enhancement of mutual understanding and economic relations at the grassroots level has thus become a key challenge. Socio-economic relations in the Asian region are more vital today than ever before.

In response to the changes detailed above, the A-STAR (Nakamura et al., 2007) consortium was established in 2006, jointly coordinated by the ATR and the National Institute of Information and Communications Technology (NICT), Japan, in cooperation with several research institutes in Asia. Current A-STAR consortium members include the Electronics and Telecommunication Research Institute (ETRI) in Korea, the Agency for Assessment and Application Technology (BPPT) in Indonesia, the National Electronics and Computer Technology Center (NECTEC) in Thailand, the Center for Development of Advanced Computing (CDAC) in India, the National Laboratory of Pattern Recognition (NLPR) in China, the National Taiwan University (NTU) in Chinese Taipei, the Institute of Information Technology (IOIT) in Vietnam, and the Institute for Infocomm Research (I^{2}R) in Singapore. Partners are still being sought in Asia for translation projects involving other languages.

The consortium was founded in order to create the basic infrastructure for spoken language communication aimed at overcoming the language barriers in the Asia-Pacific region. However, as opposed to focusing on the research and development of the speech-translation technology itself, the consortium's objective is to establish an international joint-research organization for designing formats of the bilingual corpora that are essential to advance the research and development of this technology, to design and compile basic bilingual corpora between Asian languages, and—in collaboration with research institutions working in the field in the Asia-Pacific region—to standardize interfaces and data formats to connect speech-translation modules internationally.

The consortium's activities are contracted as research by the Asia Science and Technology Co-

[2] http://verbmobil.dfki.de

[3] http://nespole!.itc.it

[4] http://www.tc-star.org

[5] http://transtac.mitre.org/

[6] http://www.ewh.ieee.org/soc/sps/stc/News/NL0701/NL0701-GALE.htm

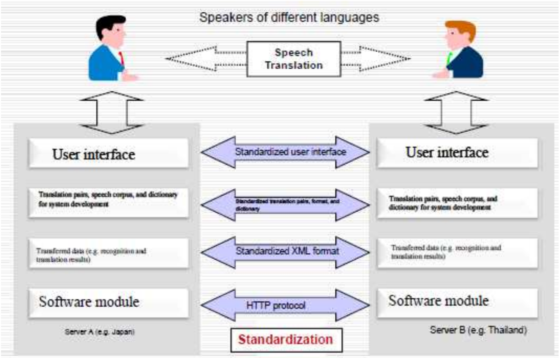
Speakers of different languages
Speech Translation
User interface
Standardized user interface
User interface
Translation pairs, speech corpus, and dictionary for system development
Standardized translation pairs, format, and dictionary
Translation pairs, speech corpus, and dictionary for system development
Transferred data (e.g. recognition and translation results)
Standardized XML format
Transferred data (e.g. recognition and translation results)
Software module
HTTP protocol
Software module
Server A (e.g. Japan)
Standardization
Server B (e.g. Thailand)

Figure 1: Standardization interface and data formats for connecting speech translation modules.

operation Promotion Strategy, which is a project undertaken by the special coordination funds for promoting science and technology. This project has further been proposed and adopted as an APEC TEL (Telecommunications and Information)[7] project. Attempts are also underway to create an expert group in the APT ASTAP (Asia-Pacific Telecommunity Standardization Program) in order to create a draft of the standardized interface and data formats for connecting speech-translation modules[8]. Figure 1 illustrates the standardized connections being considered in this project. This will standardize the interfaces and data formats of the modules comprising the speech translation architecture facilitating their connection over the Internet. It is also necessary to create common speech-recognition and speech-translation dictionaries and to compile standardized bilingual corpora. The basic communication interface will be web-based HTTP 1.1 communication, and a markup language called STML (speech translation markup language) is currently being developed as the data format for connecting applications (Kimura, 2007).

3 A-STAR Speech Technologies

In July 2009, the A-STAR consortium will be conducting large scale network-based speech-translation experiments. Eight research groups comprising A-STAR members will be participating in the conducted experiments covering nine languages that include English, Japanese, Chinese, Korean, Thai, Indonesian, Malay, Vietnamese, and Hindi.

Most of the members will be providing various speech technology servers, as described in Table 1. In total, there will be ASR engines for 8 different languages, TTS engines for 9 languages, and MT engines for (9 x 8) = 72 language combinations.

There will be no restriction on the kind of resource applied. All members will be able to train their ASR, MT, and TTS systems with any available resources corpora. The details of the specifications of each system provided by each member can be found in the reference sources listed in the last column of Table 1.

In addition to the above, NICT/ATR will be providing the 20K basic travel expression corpus (BTEC) (Kikui et al., 2006) text for demo experiments that contains tourism-related sentences similar to those that are usually found in phrase books for tourists traveling abroad. Some useful proper nouns, such the names of Asian tourist destinations, tourist attractions, and famous people, etc., are also included in the corpus. Tables 2 and 3 show the quality of direct translation for all the combination language pairs of the machine translation engines using the bilingual evaluation understudy (BLEU) (Papineni et al., 2002) scores and the metric evaluation of translation with explicit ordering (METEOR) (Banerjee and Lavie, 2005), respectively.

4 Connecting Speech-Translation Systems

All the speech-to-speech translation services are provided through the connections between client applications and STML servers. An example of the speech-translation scheme of a multi-party conversation is illustrated in Figure 2, and it is performed as follows:

- A Japanese user speaks an utterance on a client application. Thereupon, the client (User Ja) performs a speech-recognition operation using the STML server. (See 1-1. Request ASR (Ja) and 1-2. Result ASR (Ja)).

- The speech recognition results are sent to the conversation server. (See 2-1. Result ASR (Ja)). Then, the conversation server sends the recognition results ASR (Ja) that it received from the initial client (User Ja) to a different client to which it is connected.

[7] http://www.apectelwg.org

[8] http:www.aptsec.org/Program/ASTAP/

Table 1: Speech technology servers provided by A-STAR consortium members.

A-STAR Consortium Members	**Servers**			**System Description**
	ASR	**TTS**	**MT**	
NICT	English Japanese Chinese Indonesian	English Japanese Chinese Indonesian Malay	All combination language pairs (9x8)=72 MT engines	(Sakti et al., 2009)
ETRI	Korean	Korean	Korean-English	(Lee et al., 2009)
NECTEC	Thai	Thai	Thai-English	(Wutiwiwatchai et al., 2009)
BPPT	–	–	Indonesian-English	(Riza and Riandi, 2008)
CDAC	–	Hindi	Hindi-English	(S. Arora, 2009)
IOIT	Vietnamese	Vietnamese	Vietnamese-English	(Vu et al., 2009)
I^{2}R	Malay	–	Malay-English	(B. Chen, 2008)

Table 2: Translation Quality of Direct Translation Approaches (BLEU)

SRC\TRG	en	hi	id	ja	ko	ms	th	vi	zh
en	–	29.25	46.32	34.29	31.36	46.99	42.74	46.23	26.61
hi	40.77	–	32.00	26.33	24.11	32.47	31.89	32.52	20.07
id	47.01	27.49	–	32.06	30.53	77.18	40.53	41.04	26.42
ja	29.83	14.86	25.18	–	62.36	24.22	29.60	25.59	38.72
ko	27.14	13.76	23.62	63.19	–	23.00	28.01	24.75	35.20
ms	48.78	28.22	81.99	31.55	28.96	–	40.75	41.42	25.52
th	42.44	24.49	37.54	31.15	28.72	37.72	–	39.32	25.92
vi	48.87	25.43	39.10	28.92	28.95	40.44	40.54	–	23.56
zh	28.12	14.47	24.85	43.87	39.18	24.05	27.78	24.63	–

Table 3: Translation Quality of Direct Translation Approaches (METEOR)

SRC\TRG	en	hi	id	ja	ko	ms	th	vi	zh
en	–	66.41	80.85	65.94	63.25	80.66	74.20	78.49	67.22
hi	63.94	–	72.98	60.64	57.17	72.84	67.63	70.01	62.42
id	68.79	63.11	–	64.92	61.08	93.20	71.89	76.07	65.11
ja	55.15	51.97	64.80	–	83.98	65.95	63.00	64.03	75.11
ko	53.48	48.16	65.04	85.02	–	64.71	61.02	61.31	74.65
ms	70.17	64.77	94.41	63.38	60.45	–	72.55	75.99	64.45
th	64.92	59.37	72.41	63.19	59.37	72.87	–	72.08	63.10
vi	69.73	61.83	74.62	59.92	59.35	75.80	72.61	–	63.34
zh	54.24	51.03	65.20	74.06	69.87	64.08	61.61	61.74	–

(See 2-2. Result ASR (Ja) and 2-3. Result ASR (Ja)).

- The client that receives the speech recognition results from the initial client (User Ja) translates them to its own language using the STML server. (See 3-1. Request MT (Ja→Ko), 3-2. Result MT (Ja→Ko), 3-1. Request MT (Ja→Id), 3-2. Result MT (Ja→Id))
- Text-to-speech is performed using the STML server for clients that request it. (See 4-1. Request TTS (Ko) and 4-2. Result TTS (Ko))

This translation mechanism can be performed on all combinations of multi-party conversations comprising the English, Japanese, Chinese, Korean, Thai, Indonesian, Malay, Vietnamese, and Hindi languages.

5 Conclusion

This year, the first of the Asian network-based speech-to-speech translation experiments have been performed. Seven research groups comprising A-STAR consortium members will be joining in the experiments conducted in July 2009, covering eight major Asian languages. All the speech-to-speech translation engines have already been successfully implemented into web-servers that can be accessed by client applications world-wide. This implementation has realized the desired objective of a real-time, location-free speech-to-speech translation of Asian languages. Although the experiments have proved quite successful, there are still many challenges to overcome. Some examples of these would be the need for the speech-to-speech translation engines to support more Asian languages, and the further need of including the names of more Asian tourist destinations,attractions, and other famous proper nouns in the translation database. The individual components of speech-to-speech translation are also expected to be applied in a wider range of applications, including speech-information retrieval, interactive navigation, dictation, and speech summarization.

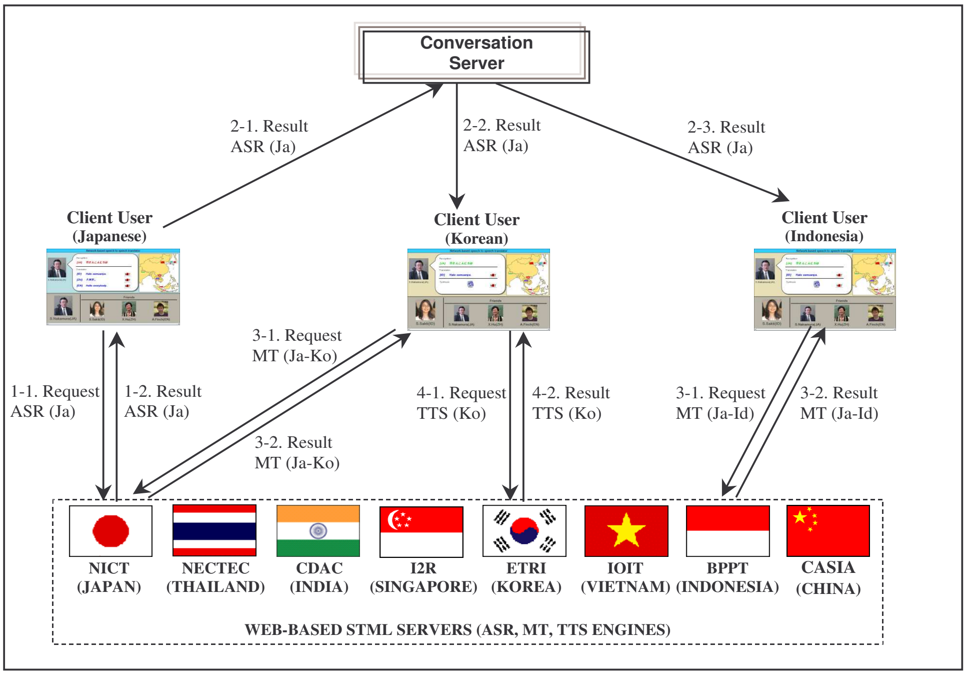

Figure 2: STML clients-servers interaction.

References

M. Zhang A. Aw H. Li B. Chen, D. Xiong. 2008. I2R multi-pass machine translation system for IWSLT 2008. In *Proc. IWSLT*, pages pp. 46–51, Hawaii, USA.

S. Banerjee and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgements. In *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72. Aan Arbor, Michigan, USA.

G. Kikui, S. Yamamoto, T. Takezawa, and E. Sumita. 2006. Comparative study on corpora for speech translation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1674–1682.

N. Kimura. 2007. Considerations for a communication interface for a multilingual speech translation platform. In *Proc. ASJ Autumn Meeting*, pages XXX–XXX, Japan.

I. Lee, J. Park, C. Kim, Y. Kim, and S. Kim. 2009. An overview of Korean-English speech-to-speech translation system. In *Proc. TCAST Workshop*, page pp. to appear, Suntec, Singapore.

S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto. 2006. The ATR multilingual speech-to-speech translation system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):365–376, March.

S. Nakamura, E. Sumita, T. Shimizu, S. Sakti, S. Sakai, J. Zhang, A. Finch, N. Kimura, and Y. Ashikari. 2007. A-STAR: Asia speech translation consortium. In *Proc. ASJ Autumn Meeting*, pages 45–46, Yamanashi, Japan.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318, Philadelphia, USA.

H. Riza and O. Riandi. 2008. Toward Asian speech translation system: Developing speech recognition and machine translation for Indonesian language. In *Proc. TCAST Workshop*, pages 30–35, Hyderabad, India.

K. Arora S. Agrawal S. Arora, R. Mathur. 2009. Development of HMM-based Hindi speech synthesis system. In *Proc. TCAST Workshop*, page pp. to appear, Suntec, Singapore.

S. Sakti, T. Vu, A. Finch, M. Paul, R. Maia, S. Sakai, T.Hayashi, N. Kimura, Y. Ashikari, E. Sumita, and S. Nakamura. 2009. Nict/atr asian spoken language translation system for multi-party travel conversation. In *Proc. TCAST Workshop*, page pp. to appear, Suntec, Singapore.

T. Vu, K. Nguyen, L. Ha, M. Luong, and S. Nakamura. 2009. Toward asian speech translation: The development of speech and text corpora for Vietnamese language. In *Proc. TCAST Workshop*, page pp. to appear, Suntec, Singapore.

C. Wutiwiwatchai, T. Supnithi, P. Porkaew, and N. Thatphithakkul. 2009. Improvement issues in English-Thai speech translation. In *Proc. TCAST Workshop*, page pp. to appear, Suntec, Singapore.