The Asian Network-based Speech-to-Speech Translation System

Sakriani Sakti [1], Noriyuki Kimura [1], Michael Paul [1], Chiori Hori [1], Eiichiro Sumita [1], Satoshi Nakamura [1],
Jun Park[2], Chai Wutiwiwatchai[3], Bo Xu[4], Hammam Riza[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] *Institute of Automation, Chinese Academy of Sciences (CASIA), China*
[5] *Agency for the Assessment and Application of Technology (BPPT), Indonesia*
[6] *Center for Development of Advance Computing (CDAC), India*
[7] *Institute of Information Technology (IOIT), Vietnam*
[8] *Institute for Infocomm Research (I²R), Singapore*

***Abstract*—This paper outlines the first Asian network-based speech-to-speech translation system developed by the Asian Speech Translation Advanced Research (A-STAR) consortium. The system was 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 Asian languages. Each A-STAR member contributes one or more of the following spoken language technologies: automatic speech recognition, machine translation, and text-to-speech through Web servers. Currently, the system has successfully covered 9 languages—namely, 8 Asian languages (Hindi, Indonesian, Japanese, Korean, Malay, Thai, Vietnamese, Chinese) and additionally, the English language. The system's domain covers about 20,000 travel expressions, including proper nouns that are names of famous places or attractions in Asian countries. In this paper, we discuss the difficulties involved in connecting various different spoken language translation systems through Web servers. We also present speech-translation results on the first A-STAR demo experiments carried out in July 2009.**

I. Introduction

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 Asian Speech Translation Advanced Research (A-STAR) consortium [1] was established in June 2006 by the National Institute of Information and Communications Technology/Advanced Telecommunications Research (NICT/ATR), Japan, with the aim of realizing network-based speech-to-speech translation systems in the Asian region, as illustrated in Fig. 1. Currently, the other seven members of the A-STAR consortium are as follows: the Electronics and Telecommunications Research Institute (ETRI) in Korea, the National Electronics and Computer Technology Center (NECTEC) in Thailand, the Institute of Automation, Chinese Academy of Sciences (CASIA) in China, the Agency for Assessment and Application Technology (BPPT) in Indonesia, the Center for Development of Advanced Computing (CDAC) in India, the Institute of Information Technology (IOIT) in Vietnam, and the Institute for Infocomm Research (I²R) in Singapore. The consortium is working collaboratively to collect Asian 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 connection between the different speech translation modules from different countries.

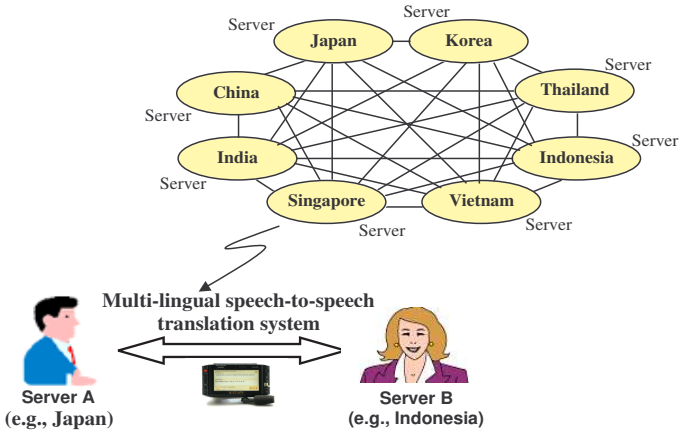

Fig. 1. Outline of the network-based speech-to-speech translation system in the Asian region.

In this paper, we outline the Asian network-based speech-to-speech translation system developed by the A-STAR consortium. The system was 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 Asian languages. Each A-STAR 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 has successfully covered 9 languages—namely, 8 Asian languages: Hindi (Hi), Indonesian (Id), Japanese (Ja), Korean (Ko), Malay (Ms), Thai (Th), Vietnamese (Vi), Chinese (Zh) and additionally, the English (En) language. The system domain covers about 20,000 travel expressions, including proper nouns–that is, the names of famous places or attractions in Asian countries.

First, a brief overview of the architecture of the A-STAR speech-to-speech translation system architecture is provided in Section II. Then, in Section III, we describe the data resources and development of the spoken language translation engines, including ASR, MT, and TTS engines. Next, the handling of the proper nouns is discussed in Section IV. The standardized data format and client application are described in Section V. We then discuss the first A-STAR demo experiments and speech-translation results in Section VI. Finally, our conclusion is presented in Section VII.

II. Overall System Architecture

Figure 2 illustrates the overall structure of our Asian network-based speech-to-speech translation system. This system is composed of the following components:

- **Spoken language technology servers**

  The spoken language technologies, including ASR, MT, and TTS engines, were provided by A-STAR members through Web servers.

- **Speech Translation Markup Language (STML) servlet**

  All data exchanges among client users and spoken language technology servers are managed through a Web service designed by NICT, the so-called STML servlet. It follows a standard protocol, namely, STML.

- **Client application**

  The client applications are implemented on a handheld mobile terminal device, which allows portable speech-to-speech translation. It was developed by NICT and supports both speech and video interaction between client users.

- **Communication server**

  A communication server, also provided by NICT, is used to relay the speech results from one user to all other users in order to enable them to perform a multiparty conversation.

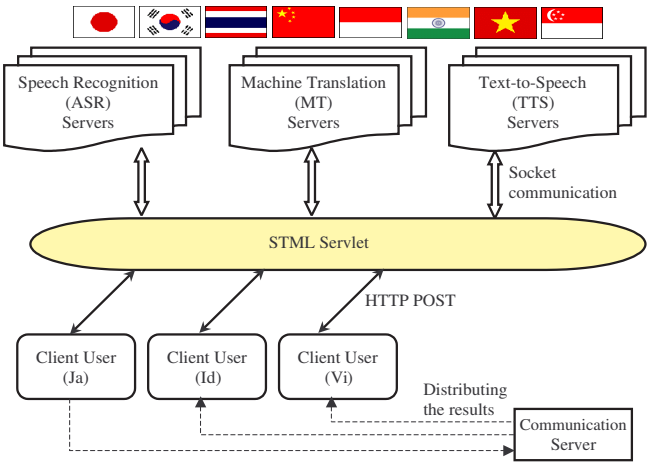

Fig. 2. Architecture of client-server interaction.

When a Japanese user utters an sentence like "*Konnichiwa*," the client application will perform a speech-recognition operation on available Japanese ASR server through the STML servlet. Then, the speech recognition results are sent to the communication server in order to be distributed to all the other clients, who are connected to each other. The clients, for example, an Indonesian user and a Vietnamese user, who receive the speech recognition results from a Japanese user, will perform a translation operation on the available Indonesian and Vietnamese MT servers through the STML servlet. In this case, the servers provide a sentence output in Indonesian, "*Selamat siang*," and one in Vietnamese, "*Xin cha'o*." Finally, the TTS servers of these languages produce the speech of the resulting sentence. This translation mechanism can be used for translating any multiparty conversation comprising any or all of the following languages—Hindi, Indonesian, Japanese, Korean, Malay, Thai, Vietnamese, Chinese, or English.

III. Spoken Language Technology Servers

Table I shows the spoken language technology servers provided by A-STAR members. In total, there will be ASR engines for 8 different languages, TTS engines for 9 languages, and MT engines for (9 x 8) = 72 different language combinations. There is no restriction on the type of resource applied. All members are allowed to train their ASR, MT, and TTS systems with any available resources corpora. The development of these spoken language technologies are described in the following section.

A. Automatic Speech Recognition

The English, Japanese, Chinese, and Indonesian ASR server engines were provided by NICT; these were trained using 202.2 hours of English travel expressions, 270.9 hours of Japanese travel expressions, 249.2 hours of Chinese travel

TABLE I
SPEECH TECHNOLOGY SERVERS PROVIDED BY A-STAR MEMBERS.

A-STAR 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	[2]
ETRI	Korean	Korean	Korean-English	[3]
NECTEC	Thai	Thai	Thai-English	[4]
BPPT	–	–	Indonesian-English	[5]
CDAC	–	Hindi	Hindi-English	[6]
IOIT	Vietnamese	Vietnamese	Vietnamese-English	[7]
I^{2}R	Malay	–	Malay-English	[8]

expressions, and 79.5 hours of Indonesian daily news [9], respectively. Parts of these systems previously belonged to the ATR multilingual speech-to-speech translation systems [10]. They were developed on the basis of hidden Markov network (HMnet) topology, which is obtained using a successive state splitting (SSS) training algorithm based on the minimum description length (MDL) optimization criterion. Details about HMnet topology based on MDL-SSS can be found in [11]. The ETRI Korean ASR engines were developed on the basis of the finite state transducer (FST), using 330 hours of speech data, including travel expressions, monologue speech, and other diverse forms of speech. The architecture of the NECTEC Thai ASR is similar to that of the JULIUS[1] speech recognizer developed by Kyoto University, Japan. It was trained using the Thai LOTUS corpus and NECTEC-ATR corpus. The IOIT Vietnamese ASR was developed on the basis of hidden Markov model (HMM), using 40.5 hours of Vietnamese radio broadcasts. The I^{2}R Malay ASR was developed on the basis of HMM, supported by the I^{2}R Abacus platform [8].

B. Machine Translation

The MT engines for all combinations of language pairs are provided on NICT translation servers. They were trained on the multilingual *Basic Travel Expressions Corpus* (BTEC)[12], which is a collection of sentences that bilingual travel experts consider useful for people going to or coming from another country [13]. The BTEC corpus used for training consists of 20K sentences for all of the involved nine languages whereby all data sets are aligned at the sentence-level.

The characteristics of the corpus languages are summarized in Table II. These languages differ largely in *word order* (subject-object-verb (SOV), subject-verb-object (SVO)), *segmentation unit* (phrase, word, syllable, none), and *morphological richness*[2] (poor, medium, rich). Concerning word segmentation, the corpora were preprocessed using language-specific word-segmentation tools for languages that do not use white-space to separate word/phrase tokens like Japanese, Korean, Thai, and Chinese. For all other languages, simple tokenization tools were applied.

[1]JULIUS,http://julius.sourceforge.jp/en_index.php

[2]The morphological classification of each language shows its relative position in these nine languages only.

TABLE II
LANGUAGE CHARACTERISTICS

Language	Word Order	Segmentation Unit	Morphological Richness
English	SVO	word	medium
Hindi	SOV	word	rich
Indonesian	SVO	word	rich
Japanese	SOV	none	rich
Korean	SOV	phrase	rich
Malay	SVO	word	rich
Thai	SVO	none	poor
Vietnamese	SVO	syllable	poor
Chinese	SVO	none	poor

TABLE III
LANGUAGE RESOURCES

BTEC		train	dev	eval
# of sen		20,000	1,000	1,000
English	voc	6,033	1,262	1,292
	len	7.2	7.1	7.2
	perpl	–	19.1	19.9
Hindi	voc	9,549	1,558	1,588
	len	7.6	7.4	7.5
	perpl	–	25.1	25.8
Indonesian	voc	6,993	1,433	1,394
	len	6.5	6.3	6.4
	perpl	–	24.4	23.4
Japanese	voc	7,002	1,407	1,408
	len	8.2	8.1	8.2
	perpl	–	15.2	15.0
Korean	voc	6,628	1,366	1,365
	len	7.8	7.7	7.8
	perpl	–	16.3	15.7
Malay	voc	7,209	1,459	1,438
	len	6.6	6.4	6.5
	perpl	–	24.7	23.5
Thai	voc	3,471	1,081	1,053
	len	7.4	7.3	7.4
	perpl	–	24.2	23.4
Vietnamese	voc	4,373	1,245	1,267
	len	8.7	8.5	8.6
	perpl	–	17.7	18.0
Chinese	voc	5,996	1,312	1,301
	len	6.6	6.4	6.5
	perpl	–	28.6	26.3

Table III summarizes the characteristics of the BTEC corpus data sets used for the training (*train*) of the SMT models, the tuning of model weights (*dev*), and the evaluation of translation quality (*eval*). Besides the number of sentences (*sen*) and the vocabulary (*voc*), the sentence length (*len*) is also given, as the average number of words per sentence. In order to get an idea how difficult the translation tasks may be, we calculated the language perplexity (*perpl*) of the respective evaluation data sets according to the language model used by the baseline system. The numbers in Table III indicate that Chinese and Hindi are the most difficult languages, followed by Malay, Indonesian, Thai, English, Vietnamese, Korean, and then Japanese.

For the training of the SMT models, standard word alignment [14] and language modeling [15] tools were used. Minimum error rate training (MERT) was used to tune the decoder's parameters, and performed on the *dev* set using the technique proposed in [14]. For the translation, a multi-stack phrase-based decoder [16] was used.

C. Text-to-Speech

As for the TTS engines, the NICT English, Japanese, and Chinese speech synthesis engines were developed on the basis of a waveform concatenation algorithm [17] in which appropriate subword units were selected from speech databases. They were trained using 16 hours of English male voices, 60 hours of Japanese female voices, and 20 hours of Chinese female voices, respectively. The NECTEC Thai speech synthesis, which is called VAJA, was also developed based on variable-length unit selection using 14 hours of Thai female voices. The NICT Indonesian/Malay, ETRI Korean, CDAC Hindi, and IOIT Vietnamese speech synthesis engines were developed using 2 hours of Indonesian female voices, 3 hours of Korean female voices, 1 hour of Hindi male voices, and 3 hours of Vietnamese male voices, respectively. Since the training data were not large enough to build unit-concatenation speech synthesis engines, they were developed on the basis of statistical parametric systems [18] in which speech waveforms are generated through parameters directly obtained from the HMMs. This system offers the ability to model different speech styles without the need for recording very large databases.

More 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 I.

IV. Handling of Proper Nouns

We have also collected additional multiparty dialog scenarios including famous proper nouns from 10 Asian countries: India, Indonesia, Japan, Korea, Malaysia, Singapore, Thailand, Vietnam, China, and the USA. The proper nouns mainly consist of city name (e.g., Kyoto in Japan, Beijing in China), tourist area (e.g., Bulkuksa in Korea, Wat Pra Kaew in Thailand), attractions (e.g., Wayang kulit in Indonesia, Kathak in India), etc.

To handle the proper nouns that never appear or appear less frequently in the training data, we enhance the language model (LM) reliability by interpolating both the class bigram of BTEC text and the new word bigram of the dialog scenario as follows:

$$\alpha P_{btec}(c_i|c_{i-1})P(w_i|c_i) + \beta P_{dialog}(w_i|w_{i-1}), \qquad (1)$$

where α represents the weight of BTEC LM, and β represents the weight of dialog LM. The optimal value of interpolation weights are estimated using a development set other than training or using the cross-validation method [19]. In this study, we estimated the interpolation weights using a development set.

V. Standardization of Data Format and Client Application

To enable communication and the exchange of data between client users and various speech language technology servers, all A-STAR members agreed to use the same STML libraries, communication data format, as well as client applications that are provided by NICT.

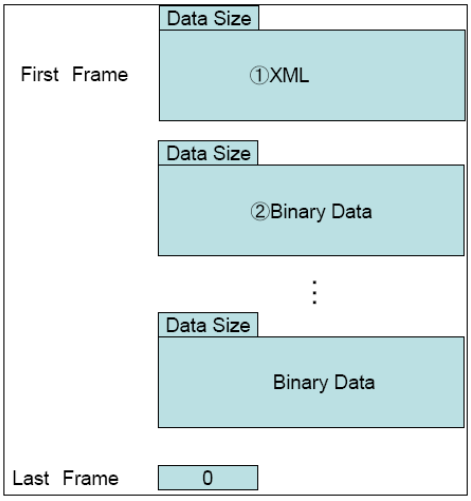

Fig. 3. Standardization of communication data format.

Fig. 4. The client application on a hand-held terminal device.

The communication data format consists of extensible markup language (XML) and binary data including a terminal frame, as described in Fig. 3. Each frame consists of a 4-byte header and actual data. The first frame is an XML header, followed by binary data from the second frame. For the terminal frame, the data size of the header is set to 0.

The input speech format is 16-kHz, 16-bit, mono-channel, and big endian / little endian. This allows the raw, distributed speech recognition (DSR), or adaptive differential pulse code modulation (ADPCM) speech format. The recognition results are N-Best with a surface form or confusion network. MT input and output strings either have a surface form or a confusion network. As for the speech synthesis, the input string has a surface form, and the output sound format is 16-kHz/8-kHz, 16-bit, mono-channel, and big endian / little endian. This allows the raw or ADPCM speech format.

The client applications are implemented on a handheld mobile terminal device (Sony VAIO-U) shown in Fig. 4, which allows portable speech-to-speech translation. The device is 150-mm wide, 95-mm high, and 32-mm thick. A unidirectional microphone is used for speech recognition in noisy

TABLE IV
TRANSLATION QUALITY (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	–

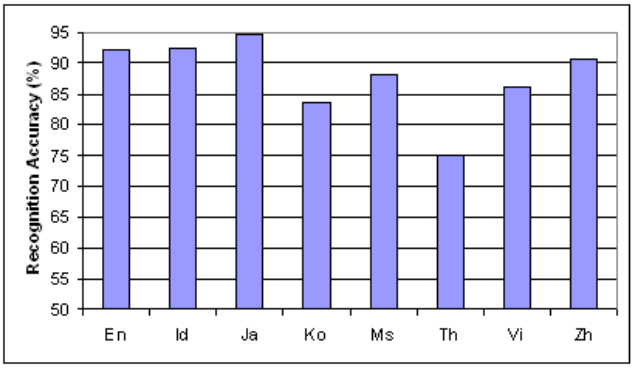

Fig. 5. Recognition accuracy rates of the English, Indonesian, Japanese, Korean, Malay, Thai, Vietnamese and Chinese ASR engines on BTEC test set.

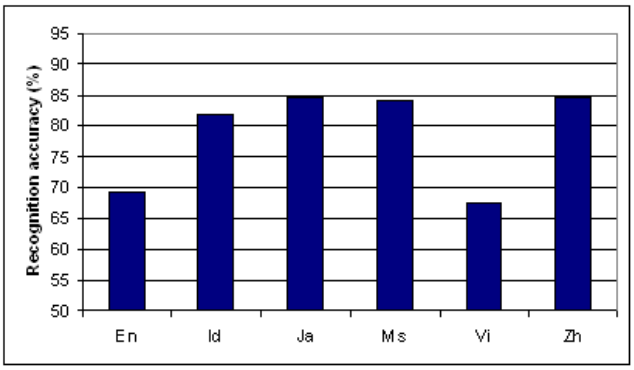

Fig. 6. Recognition accuracy rates of the English, Indonesian, Japanese, Malay, Vietnamese and Chinese ASR engines on dialog scenario test set.

environments. The system uses Microsoft Windows and a Java platform operating environment. The client application also supports an external TV conference feature that allows both speech and video interaction.

VI. EVALUATION OF TRANSLATION SERVER COMPONENTS

The first A-STAR demo experiments for connecting spoken language technologies in the Asian region were carried out in July 2009. Two types of evaluation were conducted as described in the following section.

A. Evaluation on BTEC test set

Figure 5 shows the recognition accuracy rates of the English, Indonesian, Japanese, Korean, Malay, Thai, Vietnamese, and Chinese ASR engines on the BTEC test set. The evaluation of the BTEC test set was performed offline using the randomly selected of recorded speech files from the total of 20,400 utterances spoken by 40 different speakers (20 males, 20 females) where each speaker uttered the same 510 BTEC test sentences. The best performance was achieved by the Japanese ASR, with 94.87% word accuracy.

For the automatic evaluation of translation quality, we applied the standard BLEU metrics that calculates the geometric mean of n-gram precision by the system output with respect to reference translations [20]. Scores range between 0 (worst) and 1 (best). The experimental results summarized in Table IV which show the quality of direct translation for all the 72 language pairs of the MT engines. This evaluation was performed offline on BTEC text evaluation data without the impact of ASR errors. The best performance was achieved by Malay-Indonesian MT with 81.99 BLEU score, due to high similarity of both languages in syntactic and semantic structure.

TABLE V
SUBJECTIVE EVALUATION

Target Language	Perfect (P)	Acceptable (A)	Incorrect (I)
English	67.2%	13.8%	19.0%
Indonesian	70.4 %	20.8 %	8.8%
Japanese	41.1%	22.8%	36.1%
Malay	72.7%	11.0%	16.3%
Vietnamese	72.4%	13.8%	8.8%
Chinese	50.5%	29.3%	20.2%

B. Evaluation on dialog scenario test set

In order to get an idea of how good the overall speech-translation quality when online conversation are performed, we conducted an evaluation on dialog scenario test set. It consists of 228 dialog sentences in total. Due to time limitation, we only perform the evaluation on the subset languages.

The recognition results are shown in Fig. 6. The performance drops from offline BTEC test set to online dialog scenario test set, due to different speaking style between read speech and conversational speech. Another reason is that there

are many proper nouns that are not familiar to the speakers. For example, native English speakers faced more difficulties in pronouncing some specific Asian proper nouns as compared with other Asian speakers. Hence, in such cases, the English ASR displays the largest decline.

As for the machine translation, the automatic evaluation metrics described in previous section are designed to judge the translation quality of the MT system outputs on a document level, i.e., scores are calculated on the sets of all evaluation data sentences, but not at the sentence level. In order to get an idea of how good the translation quality of a single sentence is, we also conducted a subjective evaluation where a native speaker of the target language had to assign one of the following grades to each of the evaluation sentences:

- **"P"** *perfect* translation
- **"A"** translation contains errors, but is still *acceptable*
- **"I"** translation is *incorrect*

For the subjective evaluation, we translated the English/Japanese ASR outputs of the demo scenario test set into all nine target languages. The subjective evaluation results are summarized in TableV.

VII. Conclusion

The first of the Asian network-based speech-to-speech translation experiments were performed in July 2009. Eight research groups comprising the A-STAR consortium members participated in the experiments, covering 8 Asian languages and additionally, the English language. All the speech-to-speech translation engines have already been successfully implemented into Web servers that can be accessed via client applications worldwide. This implementation has realized the desired objective of the real-time, location-free speech-to-speech translation of Asian languages. Although the experiments have proved rather successful, there are still many challenges to be overcome. Some of these are the need for speech-to-speech translation engines that can handle conversational speech as well as out-of-vocabulary words, as there are many new proper noun entries. New A-STAR partners are now being sought in Asia for translation projects involving other languages. Thus, future directions may include work and research to support other Asian spoken language technologies and the inclusion of more Asian tourist proper nouns. The individual components of speech-to-speech translation are also expected to be employed in a wider range of applications, including speech-information retrieval, interactive navigation, dictation, and speech summarization.

References

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

[2] S. Sakti, T. Vu, A. Finch, M. Paul, R. Maia, S. Sakai, T. Hayashi, N. Kimura, Y. Ashikari, E. Sumita, and S. Nakamura, "NICT/ATR asian spoken language translation system for multi-party travel conversation," in *Proc. TCAST Workshop*, Suntec, Singapore, 2009, pp. 26–30.

[3] I. Lee, J. Park, C. Kim, Y. Kim, and S. Kim, "An overview of Korean-English speech-to-speech translation system," in *Proc. TCAST Workshop*, Suntec, Singapore, 2009, pp. 6–9.

[4] C. Wutiwiwatchai, T. Supnithi, P. Porkaew, and N. Thatphithakkul, "Improvement issues in English-Thai speech translation," in *Proc. TCAST Workshop*, Suntec, Singapore, 2009, pp. 10–14.

[5] H. Riza and O. Riandi, "Toward Asian speech translation system: Developing speech recognition and machine translation for Indonesian language," in *Proc. TCAST Workshop*, Hyderabad, India, 2008, pp. 30–35.

[6] S. Arora, R. Mathur, K. Arora, and S. Agrawal, "Development of HMM-based Hindi speech synthesis system," in *Proc. TCAST Workshop*, Suntec, Singapore, 2009, pp. 31–34.

[7] T. Vu, K. Nguyen, L. Ha, M. Luong, and S. Nakamura, "Toward Asian speech translation: The development of speech and text corpora for Vietnamese language," in *Proc. TCAST Workshop*, Suntec, Singapore, 2009, pp. 15–20.

[8] B. Chen, D. Xiong, M. Zhang, A. Aw, and H. Li, "I2R multi-pass machine translation system for IWSLT 2008," in *Proc. IWSLT*, Hawaii, USA, 2008, pp. 46–51.

[9] S. Sakti, E. Kelana, H. Riza, S. Sakai, K. Markov, and S. Nakamura, "Recent progress in developing Indonesian large-vocabulary corpora and LVCSR system," in *Proc. MALINDO Workshop*, Cyberjaya-Selangor, Malaysia, 2008, pp. 40–45.

[10] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, "The ATR multilingual speech-to-speech translation system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 365–376, March 2006.

[11] T. Jitsuhiro, T. Matsui, and S. Nakamura, "Automatic generation of non-uniform HMM topologies based on the MDL criterion," *IEICE Trans. Inf. & Syst.*, vol. E87-D, no. 8, pp. 2121–2129, 2004.

[12] G. Kikui, S. Yamamoto, T. Takezawa, and E. Sumita, "Comparative study on corpora for speech translation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1674–1682, 2006.

[13] M. Paul, H. Okuma, H. Yamamoto, E. Sumita, S. Matsuda, T. Shimizu, and S. Nakamura, "Multilingual mobile-phone translation services for world travelers," in *Proc. COLING, Companion Volume*, Manchester, UK, 2008, pp. 165–168.

[14] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proc. ACL*, Sapporo, Japan, 2003, p. 160167.

[15] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. ICSLP*, Denver, USA, 2002, pp. 901–904.

[16] A. Finch, E. Denoual, H. Okuma, M. Paul, H. Yamamoto, K. Yasuda, R. Zhang, and E. Sumita, "The NICT/ATR speech translation system for IWSLT 2007," in *Proc. IWSLT*, Trento, Italy, 2007, pp. 103–110.

[17] N. Campbell and A. Black, "Prosody and the selection of source units for concatenative synthesis," in *Progress in speech synthesis*, J. van Santen, R. Sproat, J. Olive, and J. Hirschberg, Eds. Springer-Verlag, 1996, pp. 279–282.

[18] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, Istanbul, Turkey, 2000, pp. 1315–1318.

[19] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*. New Jersay, USA: Prentice Hall, 2001.

[20] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. ACL*, Philadelphia, USA, 2002, pp. 311–318.