NICT/ATR Asian Spoken Language Translation System for Multi-Party Travel Conversation

Sakriani Sakti, Thang Tat Vu, Andrew Finch, Michael Paul, Ranniery Maia, Shinsuke Sakai, Teruaki Hayashi, Shigeki Matsuda, Noriyuki Kimura, Yutaka Ashikari, Eiichiro Sumita, Satoshi Nakamura
NICT Spoken Language Communication Research Group *
2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan
{sakriani.sakti,thang.vu,andrew.finch,michael.paul,ranniery.maia}@nict.go.jp
{shinsuke.sakai,teruaki.hayashi,shigeki.matsuda,noriyuki.kimura}@nict.go.jp
{yutaka.ashikari,eiichiro.sumita,satoshi.nakamura}@nict.go.jp

Abstract

This paper presents the recent advances in the Asian spoken language translation system developed by the National Institute of Information and Communications Technology/Advanced Telecommunications Research Institute International (NICT/ATR). The system was designed to translate the common spoken utterances of travel conversation from a certain source language into multi-target languages in order to facilitate a multi-party travel conversation between people speaking different Asian languages. All the system engines, including those of speech recognition, machine translation, and speech synthesis, were developed using the corpus-based approach, which statistically studied from the collection of speech and language data. Currently, all the speech-to-speech translation engines have been successfully implemented into web-servers that can be accessed by client applications worldwide. Thus, the system realizes the objective of real-time, location-free speech-to-speech translation.

1 Introduction

With the increase in globalization, international tourism, and international business, the issue of communication between people not sharing a common language has gained importance. Humankind has long dreamt of realizing a speech translation technology that will be able to break down the language barrier and thus enable instant cross-lingual communication. Many researchers have been working in the area of speech recognition, machine translation, and speech synthesis for almost five decades. With the ongoing advances in spoken language processing technology, the dream of realizing a multilingual speech translation system has become more feasible.

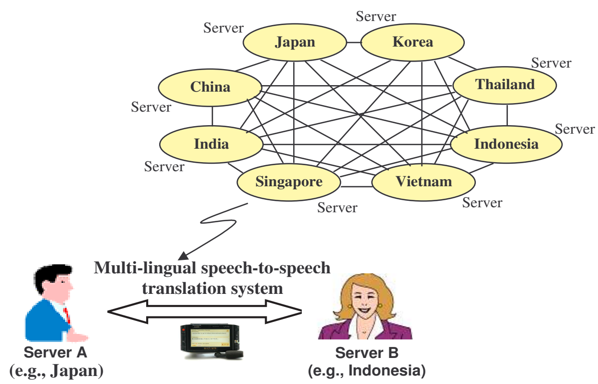

Figure 1: Outline of future speech-technology services connecting each area in the Asian region through a network.

In Asia, the speech translation advanced research (A-STAR) (Nakamura et al., 2007) consortium was formed in 2006. This consortium comprises Asian speech and natural language researchers that working together in order to advance the development of state-of-the-art multilingual man-machine interfaces in the Asian region. This basic infrastructure helps to accelerate the development of large-scale spoken language corpora in Asia and also to facilitate the development of multilingual speech translation systems including automatic speech recognition, machine translation, and speech synthesis systems. These fun-

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

damental technologies are expected to be applicable to the human-machine interfaces of various telecommunication devices and services connecting Asian countries through a network using standardized communication protocols, as outlined in Fig. 1.

In this paper, we outline our contributions toward the development of a spoken language translation system catering to Asian languages within the A-STAR project. The system is designed to translate the common spoken utterances of tourists from a certain source language into multi-target languages, so that a multi-party travel conversation between tourists speaking different Asian languages can take place effectively. Currently, the system has successfully covered four Asian languages, including Japanese (Ja), Chinese (Zh), Indonesian (Id), and Vietnamese (Vi). Additionally, we also include English (En) language into the system, since it is the most widely spoken second language in the Asia region. All the system engines, including the speech recognition, machine translation, and speech synthesis system engines, were developed using the corpus-based approach, which statistically studied from the collection of speech and language data.

We first briefly provide an overview of the NICT/ATR multilingual speech-to-speech translation system architecture in Section 2. We then describe the work and experiments related to the speech recognition engine (Section 3), machine translation engine (Section 4), and speech synthesis engine (Section 5). The integration of these component technologies into a web-based speech translation system is described in Section 6. Finally, we present our summary in Section 7.

2 Overall System Architecture

Figure 2 provides an overview of the NICT/ATR multilingual speech-to-speech translation system architecture. The system has been designed for the purpose of translating spoken utterances of a certain source language into multi-target languages.

The NICT/ATR multilingual speech-to-speech translation system basically consists of three parts, namely an automatic speech recognition (ASR) engine, a machine translation (MT) engine, and a speech synthesis / text-to-speech (TTS) engine. When a person utters a Japanese sentence like "*Konnichiwa*," the system attempts to recognize the input speech utterance through a Japanese

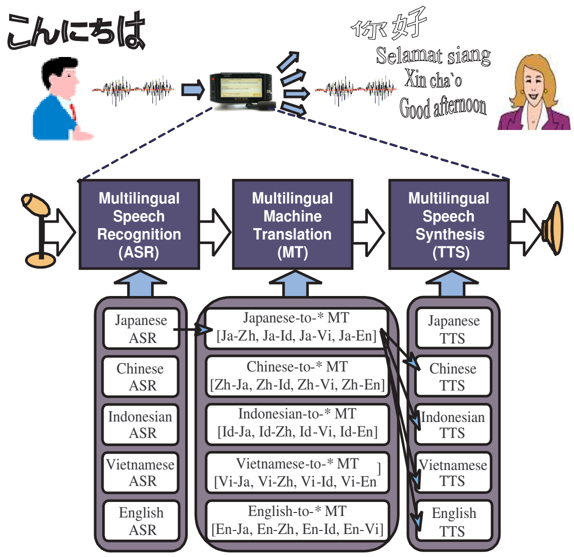

Figure 2: Overview of the multilingual speech-to-speech translation system architecture.

speech recognizer. Following this, the resulting Japanese text sentence is translated into multi-target language sentences—like Chinese, Indonesian, Vietnamese, and English—by a Japanese-to-* machine translator. In this particular case, the machine gives a Chinese sentence output "*Ni Hao*," an Indonesian sentence output "*Selamat siang*," a Vietnamese sentence output "*Xin cha'o*," and an English sentence output "*Good afternoon*." Finally, all the synthesizers of these languages produce the spoken output of the resulting sentence. This translation mechanism can be used for translating any multi-party conversation comprising any or all of the Japanese, Chinese, Indonesian, Vietnamese, or English languages.

The details of the development of each component of the speech recognition, machine translation, and speech synthesis systems are described in the following section.

3 Speech Recognition Engine

3.1 Data Corpora

The training data of the Japanese, Chinese, Indonesian, Vietnamese, and English speech recognition engines are summarized in Table 1. The Japanese, English, and Chinese corpora were part of the ATR basic travel expression corpus (BTEC), while the Indonesian and Vietnamese corpora were developed by Indonesian and Vietnamese re-

Table 1: Training data of Japanese (Ja), Chinese (Ch), Indonesian (Id), Vietnamese (Vi), and English (En) speech recognition engines.

Lang	# Phn	# Acc	# Spkrs (M,F)	# Utts	# Hrs	Domain
Ja	26	No accent	4200 (1600, 2600)	172,674	270.9	Travel expression
Zh	85	4 (BJ, SH, CT, TW)	536 (268, 268)	207,257	249.2	Travel expression
Id	33	4 (ST, JV, SN, BT)	400 (200, 200)	84,000	79.5	Daily news
Vi	45	4 No accent	30 (15, 15)	23,424	40.5	Radio broadcast
En	44	3 (US, BRT, AUS)	532 (266, 266)	207,724	202.0	Travel expression

search institutions, respectively.

The Indonesian corpus was developed by the R&D Division of PT Telekomunikasi Indonesia (R&D TELKOM) in collaboration with ATR as a continuation of the APT (Asia Pacific Telecommunity) project (Sakti et al., 2004). This corpus contains phonetically-balanced sentences culled from the leading newspaper and magazine in Indonesia, "KOMPAS" and "TEMPO," respectively.

The Vietnamese corpus was developed by the Institute of Information Technology (IOIT), Vietnam, under the A-STAR project. The speech was recorded from "The Voice of Vietnem" (VOV) radio programs.

In total, the data corpora consist of 4200 Japanese speakers, 536 Chinese speakers, 400 Indonesian speakers, 30 Vietnamese speakers, and 532 English speakers. Some data also covers several accents; for example, the Chinese corpus includes the following accents: Beijing (BJ), Shanghai (SH), Cantonese (CT), and Taiwanese (TW). The Indonesian corpus covers these accents: Java (JV), Sundanese (SN), and Batak (BT), and additional Standard Indonesian (ST), while the English corpus includes the following: United States (US), British (BRT), and Australian (AUS) accents.

3.2 Model Training and Evaluations

Each speech recognition engine—Japanese, Chinese, Indonesian, Vietnamese, and English—was trained separately. The experiments pertaining to all the above mentioned languages were conducted using the following feature extraction parameters: sampling frequency of 16 kHz, frame length of a 20-ms Hamming window, frame shift of 10 ms, and 25 dimensional MFCC features (12-order MFCC, Δ MFCC, and Δ log power).

Three states were used as the initial hidden Markov model (HMM) for each phoneme. A shared state HMnet topology was then obtained using a successive state splitting (SSS) training algorithm based on the minimum description length

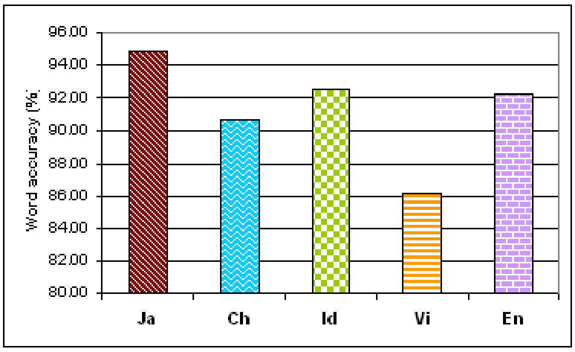

Figure 3: Recognition accuracy rates of the Japanese, Chinese, Indonesian, Vietnamese and English speech recognition engines on the ATR-BTEC test set.

(MDL) optimization criterion (Jitsuhiro et al., 2004). The composite multi-class bigram and trigram language models were trained using the 20K of the ATR-BTEC text sentences.

The system was tested on the ATR-BTEC test set. On an average, each language comprised about 40 speakers (20 males, 22 females), where each speaker uttered the same 510 BTEC sentences. The optimum performances of the speech recognition engines of all the languages are shown in Figure 3.

4 Machine Translation Engine

4.1 Data Corpora

The ATR-BTEC text data has functioned as the primary source for developing broad-coverage speech translation systems. The sentences forming this text data were collected by bilingual travel experts from among Japanese/English sentence pairs in travel domain "phrasebooks." The ATR-BTEC text data has also been translated into 18 different languages that include French, German, Italian, Chinese, Korean, Indonesian, and Vietnamese. The training set of each language corpus consists of 20,000 sentences while the test set comprises 510 sentences, with 16 references per sentence.

4.2 Model Training and Evaluations

The phrase-based statistical machine translation systems of the Indonesian-Japanese and Indonesian-English corpora were trained using the 20K sentence pairs of ATR-BTEC text data described above. Monolingual features—like the language model probability—were trained on the 20K monolingual text corpus of the target language.

For decoding, a multi-stack phrase-based SMT decoder called CleopATRa (Finch et al., 2007) was used. The translation quality was evaluated using the standard automatic evaluation metrics described in Table 2. The average of the respective BLEU and METEOR scores obtained for all the MT engines are listed in Tables 3 and 4, respectively.

More details about the NICT/ATR speech translation system can be found in these papers (Paul et al., 2008; Shimizu et al., 2008; Sumita et al., 2007).

Table 2: Automatic Evaluation Metrics

BLEU:	The geometric mean of the n-gram precision of the system output with respect to reference translations. Scores range between 0 (worst) and 1 (best) (Papineni et al., 2002)
METEOR:	A metric that calculates unigram overlaps between translation and reference texts while taking into account various levels of matches (*exact*, *stem*, *synonym*). Scores range between 0 (worst) and 1 (best) (Banerjee and Lavie, 2005)

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

SRC\TRG	ja	zh	id	vi	en
ja	–	38.72	25.18	25.59	29.83
zh	43.87	–	24.85	24.63	28.12
id	32.06	26.42	–	41.04	47.01
vi	28.92	23.56	39.10	–	48.87
en	34.29	26.61	46.23	46.23	–

Table 4: Translation Quality of Direct Translation Approaches (METEOR)

SRC\TRG	ja	zh	id	vi	en
ja	–	75.11	64.80	64.03	55.15
zh	74.06	–	65.20	61.74	54.24
id	64.92	65.11	–	76.07	68.79
vi	59.92	63.34	74.62	–	69.73
en	65.94	67.22	80.85	78.49	–

5 Speech Synthesis Engine

5.1 Data Corpora

There are large corpora (Japanese, Chinese, and English) and small corpora (Indonesian and Vietnamese) that are used for the purpose of training the speech synthesis engines that were developed by NICT/ATR. They are consist as follows:

- 60 hours of Japanese female voices
- 16 hours of English male voices
- 20 hours of Chinese female voices
- 2 hours of Indonesian female voices
- 3 hours of Vietnamese male voices

5.2 Model Training and Evaluations

The development of the Japanese, Chinese, and English speech synthesis engines was based on a waveform concatenation algorithm in which appropriate subword units were selected from speech databases (Campbell and Black, 1996). Through this technique, the system could synthesize a high quality of speech.

For the Indonesian and Vietnamese languages, the training data was not large enough to build unit-concatenation speech synthesis engines. In this case, we developed statistical parametric speech synthesis systems based on the hidden Markov models (HMMs) in which speech waveforms are generated through parameters directly obtained from the HMMs (Tokuda et al., 2000). This system offers the ability to model different speech styles without the need for recording very large databases.

Here, both the Indonesian and Vietnamese statistical parametric speech synthesis systems used five state left-to-right HMMs for each phoneme-sized speech unit. These context-dependent HMMs were trained using the full contextual labels and concatenated feature vectors of extracted F_0 and mel-cepstrum parameters. The distributions for the excitation (pitch) parameter, spectral parameter, and the state duration were clustered independently using a decision-tree based context clustering technique.

The Speech waveform was synthesized using only simple excitation and the MLSA (mel-log spectrum approximation) filter (Tokuda et al., 1995). It has been observed that the Indonesian and Vietnamese speech synthesis engines are also able to synthesize speech that resembles the speaker's speech in the database. The speaking rate of the synthesized version is also similar to that of the natural speech case. Through informal listening tests, we have found that the synthesized speech still presents the buzziness that is characteristic of the simple excitation model. However, by and large, the prosody is good and the speech sounds smooth and stable.

6 Web-based Asian Speech Translation System

All the above spoken language translation technologies—speech recognition, machine translation, and speech synthesis engines catering to the Japanese, Chinese, Indonesian, Vietnamese, and English languages—have been successfully applied into web-servers that can be accessed by client applications. The client applications are implemented on a handheld mobile terminal device, thus realizing the objective of real-time, location-free speech-to-speech translation.

Currently, this NICT/ATR speech translation system has been connected to speech translation servers provided by other A-STAR consortium members, including ETRI (Korea), NECTEC (Thailand), BPPT (Indonesia), CDAC (India), IOIT (Vietnam), and I2R (Singapore). Therefore, all terminal devices can be connected, not only to any internal speech translation server, but also to any external speech translation server in the Asian region provided by others A-STAR members. Thus, the final speech translation system can translate not just between the Japanese, Chinese, Indonesian, Vietnamese, and English languages, but also between other major Asian languages.

7 Summary

In this paper, we have presented the NICT/ATR research activities carried out in accordance with the A-STAR consortium and conducted toward developing a multilingual speech-to-speech translation system catering to Asian languages. This system can translate source language spoken utterances into multi-target languages, so that a multi-party travel conversation between different Asian languages can take place. Currently, all the speech-to-speech translation engines have been successfully implemented into web-servers that can be accessed by client applications worldwide. The system has also been connected to other speech recognition, machine translation, and speech synthesis servers provided by other A-STAR consortium members. Thus, it realizes the objective of real-time, location-free speech-to-speech translation not only between the Japanese, Chinese, Indonesian, Vietnamese, and English languages, but also between other major Asian languages.

References

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.

N. Campbell and A. Black. 1996. Prosody and the selection of source units for concatenative synthesis. In J. van Santen, R. Sproat, J. Olive, and J. Hirschberg, editors, *Progress in speech synthesis*, pages 279–282. Springer Verlag.

A. Finch, E. Denoual, H. Okuma, M. Paul, H. Yamamoto, K. Yasuda, R. Zhang, and E. Sumita. 2007. The nict/atr speech translation system for iwslt 2007. In *Proc. IWSLT*, pages 103–110, Trento, Italy.

T. Jitsuhiro, T. Matsui, and S. Nakamura. 2004. Automatic generation of non-uniform HMM topologies based on the MDL criterion. *IEICE Trans. Inf. & Syst.*, E87-D(8):2121–2129.

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.

M. Paul, H. Okuma, H. Yamamoto, E. Sumita, S. Matsuda, T. Shimizu, and S. Nakamura. 2008. Multilingual mobile-phone translation services for world travelers. In *Proc. Coling 2008*, pages 21–24, Manchester, UK.

S. Sakti, P. Hutagaol, A. Arman, and S. Nakamura. 2004. Indonesian speech recognition for hearing and speaking impaired people. In *Proc. ICSLP*, pages 1037–1040, Jeju, Korea.

T. Shimizu, Y. Ashikari, E. Sumita, J. Zhang, and S. Nakamura. 2008. NICT/ATR Chinese-Japanese-English speech-to-speech translation system. *Tsinghua Science and Technology*, 13(4):540–544.

E. Sumita, T. Shimizu, and S. Nakamura. 2007. NICT-ATR speech-to-speech translation system. In *Proc. ACL*, pages 25–28, Prague, Czech Republic.

K. Tokuda, T. Kobayashi, and S. Imai. 1995. Adaptive cepstral analysis of speech. *IEEE Trans. Speech and Audio Processing*, 3(6):481–489.

K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. ICASSP*, pages 1315–1318, Istanbul, Turkey.