Development of Indonesian Spoken Language Technologies for Multilingual Speech-to-Speech Translation System

Sakriani Sakti, Michael Paul, Ranniery Maia, Shinsuke Sakai, 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,michael.paul,ranniery.maia,shinsuke.sakai,noriyuki.kimura}@nict.go.jp
{yutaka.ashikari,eiichiro.sumita,satoshi.nakamura}@nict.go.jp

Abstract

In this paper, we present the recent advancements in the development of Indonesian spoken language technologies towards a multilingual speech-to-speech translation system. These comprise the development of speech recognition, language translation, and speech synthesis systems. All the system models are based on a corpus-based approach and were statistically trained using a collection of speech and language data. The experimental results reveal that the Indonesian speech recognition system could achieve 92.47% word accuracy on travel domain, while the BLEU (bilingual evaluation understudy) machine translation score was about 0.57 and 0.6 for Indonesian-Japanese and Indonesian-English translations, respectively. The quality of Indonesian synthesized speech was also smooth and perfectly intelligible.

1. Introduction

An automatic spoken language translation system has been a long-standing dream for people all over the world, as such a system would enable instant cross-lingual communication and overcome prevalent language barriers. The use of mechanical dictionaries was first suggested in the seventeenth century, but it was not until the twentieth century that the first patent for "translating machines" was successfully applied [1]. In Japan, a translating machine project was started at the ATR (Advanced Telecommunications Research Institute International) in the 1980s as a national project. As a result of this research, the component technologies—speech recognition, language translation, and speech synthesis—progressed tremendously.

At present, NICT/ATR (National Institute of Information and Communications Technology/Advanced Telecommunications Research Institute International) has developed multilingual translation systems for providing both text-to-text and speech-to-speech services. Both systems have been successfully implemented into mobile phone applications in the form of the first commercial speech translation service in the world. The multilingual text-to-text translation system covers 18 of the world's major languages and is currently able to provide translated travel conversations in 306 (=18 x 17) language pairs, while the speech-to-speech translation system covers the Japanese and English languages for real environments [2]. Additionally, NICT/ATR has developed a Chinese-Japanese-English speech-to-speech translation system on a hand-held terminal device [3]. This system is now ready to be made available to world travelers.

In this paper, we outline our contribution towards developing Indonesian spoken language technologies for the NICT/ATR multilingual speech-to-speech translation system. We have developed an Indonesian speech recognition system, an Indonesian speech synthesis system, as well as Indonesian-English and Indonesian-Japanese language translation systems. All the above system models are based on a corpus-based approach and were statistically trained using a collection of speech and language data.

The rest of this paper is organized as follows. Section 2 provides an overview of the NICT/ATR multilingual speech-to-speech translation system architecture. Section 3 describes the development of Indonesian speech technologies, including speech recognition, language translation, and speech synthesis. Section 4 describes the hand-held speech-to-speech translation terminal devices. Section 5 describes the outline of the plan devised for developing speech translation systems for other Asian languages. Finally, Section 6 presents the summary.

*Spoken Language Communication Research Group of NICT was previously belonging to ATR Spoken Language Communication Research Laboratories, Japan

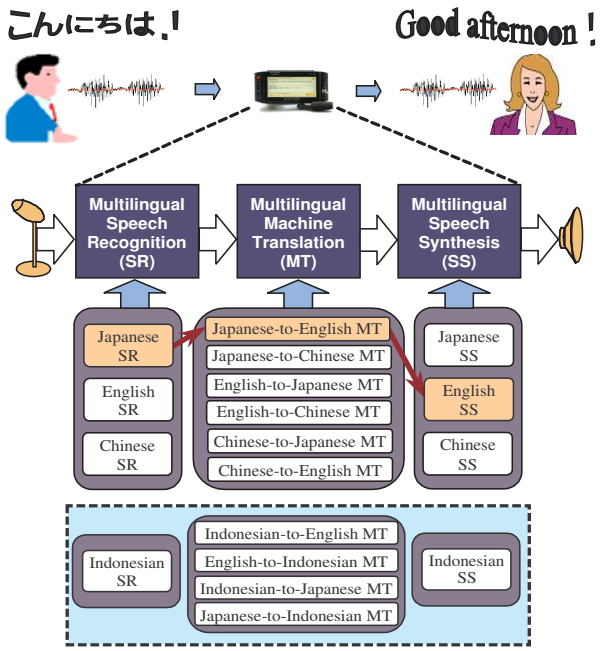

Figure 1. Overview of the multilingual speech-to-speech translation system architecture.

2. Overall System Architecture

An overview of the NICT/ATR multilingual speech-to-speech translation system architecture is illustrated in Figure 1. The system is designed in order to translate the spoken utterance of a certain source language into a target language. It basically comprises three parts, namely, a speech recognition engine, language translation engine, and speech synthesis engine. When a person utters a Japanese sentence like "*Konnichiwa*", the system attempts to recognize the input speech utterance using a Japanese speech recognizer. After that, the resulting Japanese text sentence is translated into a target language sentence—for example, English—by a Japanese-to-English machine translator. In this case, it provides the translated English sentence output "*Good afternoon*". Finally, the English synthesizer is used to produce the spoken output of the resulting English sentence. This translation mechanism can be used for any language pair from among Japanese, English, and Chinese.

In order to translate Indonesian spoken utterances into/from other languages, we need to add several components related to the Indonesian speech technologies; these are shown within the dotted box in Figure 1. The development of these components will be described in detail in the following section.

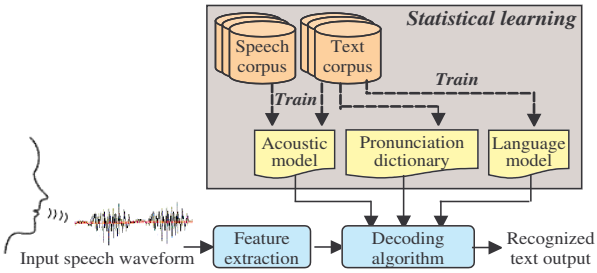

Figure 2. Basic framework of the speech recognition engine.

3. System Components

3.1. Speech Recognition Engine

The experiments 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).

The basic framework of the speech recognition engine is illustrated in Figure 2. The training involved the development of an acoustic model, a pronunciation dictionary, and a language model in the manner detailed below:

1. **Acoustic model**

   The acoustic model was trained using the multi-speaker clean speech data of both the daily news and telephone application tasks. The corpora were developed by the R&D Division of PT Telekomunikasi Indonesia (R&D TELKOM) in collaboration with the ATR as a continuation of the APT (Asia Pacific Telecommunity) project [4].

   There were a total of 400 speakers (200 males and 200 females). The speech of the speakers was characterized by four different speaking styles, which included the standard unaccented Indonesian and three main native language accents (Batak, Java, and Sunda). Each speaker uttered 210 sentences, resulting in a total of 84,000 speech utterances or about 80 hours of speech.

   Segmented utterances according to labels are usually used as a starting point in speech recognition systems for training speech models. Automatic segmentation is mostly used since it is efficient and less time consuming. It is basically produced by forced alignment given the transcriptions. However, in this first stage, a proper Indonesian acoustic model is not available yet. In this case, we solve this problem by developing initial Indonesian phoneme-based acoustic model using the English-Indonesian cross language approach [5].

Using the resulting segmented utterance, we trained the acoustic model. A hidden Markov model (HMM) is typically employed to represent the acoustic model. Three states were used as the initial 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 (MDL) optimization criterion [6]. The resulting context-dependent triphone had 1,277 states in total with 15 Gaussian mixture components per state.

2. **Pronunciation dictionary**

   The pronunciation dictionary, which is owned by R&D TELKOM, was derived from the daily news and telephone application text corpus. It consists of about 40,000 words in total, including 30,000 original Indonesian words plus 8000 person and place names and 2000 foreign words. The pronunciation of all these words was manually developed by Indonesian linguists. On the basis of these pronunciations, we then included additional words derived from travel expression sentences.

3. **Language model**

   Word bigram and trigram language models were trained using the 160K sentences contained in the ATR basic travel expression corpus (BTEC) [7] train set, yielding a trigram perplexity of 67.0 and an out-of-vocabulary (OOV) rate of 0.78% on the 510 sentences of the BTEC test set.

The performance of the system was tested on the ATR-BTEC data. There were 42 speakers (20 males, 22 females) and each speaker uttered the same 510 BTEC sentences, resulting in a total of 21,420 utterances (23.4 hours of speech). The Indonesian speech recognition system achieved a performance of 92.47% word accuracy at RTF = 0.97.

3.2. Language Translation Engine

The basic framework of the machine translation engine is illustrated in Figure 3. The training involved the development of translation and language models of the target language as follows:

1. **Translation model**

   Phrase-based translation models were trained for translating Indonesian-to-Japanese, Japanese-to-Indonesian, Indonesian-to-English, and English-to-Indonesian, using 160K of the ATR-BTEC sentence pairs (with about 20,000 unique words).

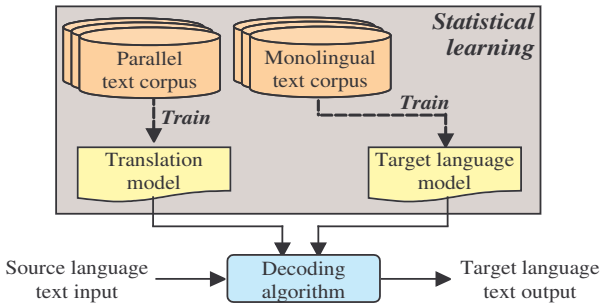

Figure 3. Basic framework of the machine translation engine.

Table 1. Results of the evaluation of BLEU and METEOR scores by SMT engines.

Language pair	BLEU (%)	METEOR (%)
Indonesian-to-Japanese	57.24	69.00
Japanese-to-Indonesian	40.59	62.33
Indonesian-to-English	59.69	75.83
English-to-Indonesian	48.35	66.82

2. **Target language model**

   A trigram and a 5-gram language models were trained for each target language on the 160K monolingual text corpora. The trigram language model was used to tune the system parameters using MERT (minimum error rate training) [8] technique, while the 5-gram language model was used during decoding.

For decoding, a multi-stack phrase-based SMT decoder called CleopATRa [9] was used. The quality of the Indonesian-Japanese and Indonesian-English SMT engines was evaluated using an evaluation data set of 510 sentences for each language, with 16 references per sentence. Table 1 shows the translation results for both Indonesian-Japanese and Indonesian-English SMT engines using the bilingual evaluation understudy (BLEU) [10] scores and the metric evaluation of translation with explicit ordering (METEOR) [11].

More details about NICT/ATR speech translation system can be found in [2, 3, 12].

3.3. Speech Synthesis Engine

The basic framework of the HMM-based speech synthesis engine is illustrated in Figure 5. The training involved the development of context-dependent HMM.

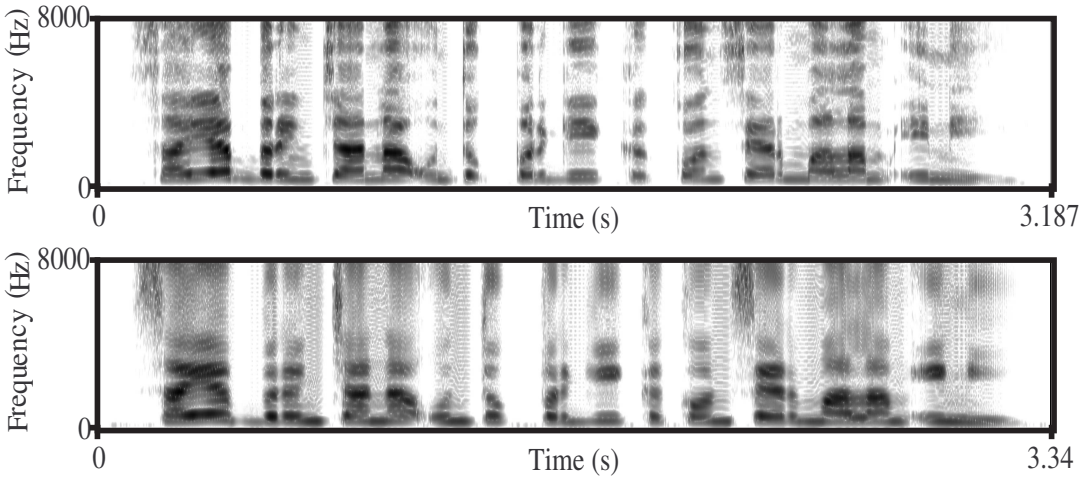

Figure 4. Spectrograms of both natural speech (top) and synthesized speech (bottom) for an utterance "*Saya berencana untuk pergi ke konser malam ini*" (meaning "*I plan on going to the concert this evening*").

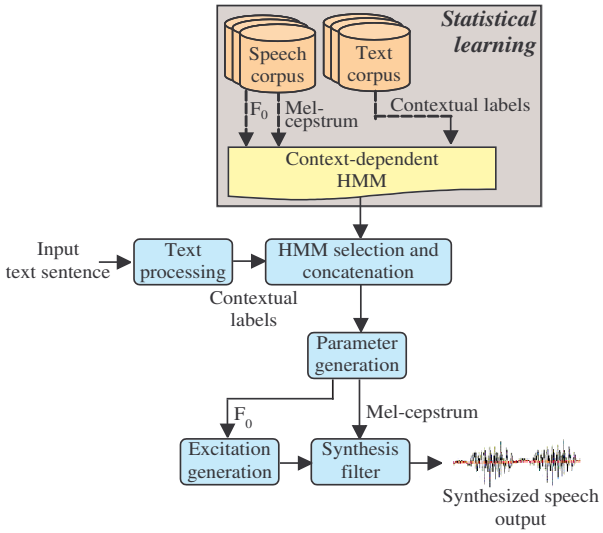

Figure 5. Basic framework of the HMM-based speech synthesis engine.

Two hours of single-speaker phonetically-balanced speech data was used as the training data. It consisted of a total of 2,012 sentences. The speech signals were sampled at a rate of 16 kHz and windowed by a 25-ms Hamming window. Then, both excitation and spectrum parameters were extracted from the speech database at a frame of every 5 ms. The excitation feature vector (pitch) consisted of log F_0 and its dynamic parameters (delta and acceleration). The spectral feature vector consisted of 25 mel-cepstral coefficients [13], including the zeroth coefficient, and their dynamic parameters (delta and acceleration).

The full contextual label was generated from a phonetic transcription using text processing tools, and it contained only the phoneme identity and its positional information features, which included the following:

- Phoneme level:
  - {second preceding, preceding, current, succeeding, second succeeding} phoneme,
  - position of current phoneme in the current word (forward and backward);
- Word level:
  - number of phonemes in {preceding, current, succeeding} word,
  - position of current word in the current utterance (forward and backward);
- Utterance level:
  - number of words in the utterance,
  - utterance types: declarative, interrogative or imperative sentence.

Five state left-to-right HMMs were used, where each HMM corresponded to a phoneme-sized speech unit. These context-dependent HMMs were trained using the full contextual labels and the concatenated feature vectors of extracted F_0 and mel-cepstrum parameters. The mel-cepstrum feature vectors were modelled by continuous probability

distribution while the F_0 feature vectors were modelled by multi-spaced probability distribution (including a discrete voiced/unvoiced symbol and one-dimensional continuous log F_0 values). The state durations of each HMM were modelled by n-dimensional Gaussians, where the dimension was equal to the number of HMM states.

The distributions for the excitation (pitch) parameter, spectral parameter, and state duration were clustered independently by using a decision-tree-based context clustering technique. By applying 1250 phonetic and positional questions, the resulting trees for spectrum, pitch, and duration models had 2,409 , 4,245, and 961 leaves, respectively.

The speech waveform was synthesized using only simple excitation and the MLSA (mel-log spectrum approximation) filter [13]. Figure 4 shows an example of the spectrogram comparisons of both natural speech (top) and synthesized speech (bottom) for the utterance "*Saya berencana untuk pergi ke konser malam ini*" (meaning "*I plan on going to the concert this evening*") which is part of the training data. It is observed that the system is able to synthesize the 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, although we have found through informal listening tests that the synthesized speech still carries the characteristic buzziness of the simple excitation model. However, the prosody is by and large good and the speech sounds smooth and stable.

4. Hand-held Terminal Devices

All the above spoken language translation technologies are integrated into a hand-held mobile terminal device, which is shown in Figure 6.

Figure 6. A hand-held terminal device of the NICT/ATR multilingual speech-to-speech translation system.

The device is 150-mm wide, 95-mm high, and 32-mm thick. A uni-directional microphone is used for speech recognition in noisy environments. In each translation process, the interface device shows both the speech recognition result and language translation result parts.

5. Towards Asian Speech Translation

Part of this project was carried out in accordance with the "Asia speech translation (A-STAR) Consortium" [14]. The goal of the project is to advance the development of multilingual man-machine interfaces, particularly the multilingual speech translation systems, in the Asian region. Thus, the final speech translation system is expected to include not only Japanese, English, Chinese, and Indonesian languages, but also other Asian languages.

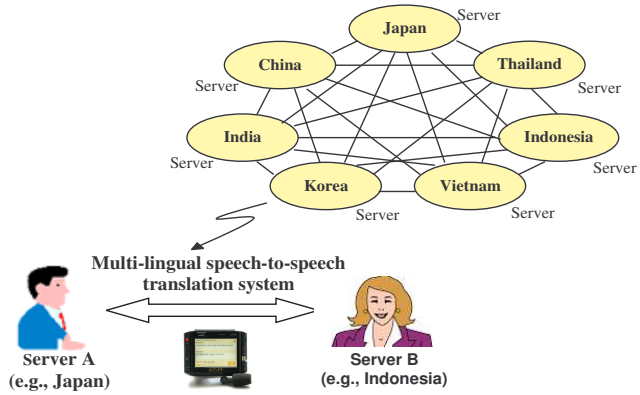

Figure 7. Outline of the Asian speech-translation services connecting each area in the Asian region through a network.

All the terminal devices can be connected to any internal or external speech-to-speech translation server in the Asian region through a machine translation markup language (STML). An outline of the Asian speech-translation services is provided in Figure 7.

These fundamental technologies are expected to be applicable to the human-machine interfaces of various telecommunication devices and services connecting Asian countries. The improvements in borderless communication made in the Asian region are expected to benefit many areas, including tourism, business, education, and social life.

6. Summary

We have presented the NICT/ATR research activities in developing a multilingual speech-to-speech translation sys-

tem which translates source language spoken utterances into the target language. We have also included the details of the development of additional components related to Indonesian spoken language technologies. These include the development of an Indonesian speech recognizer, an Indonesian speech synthesizer, and Indonesian-Japanese and Indonesian-English machine translators. Part of this project was carried out in accordance with the A-STAR consortium in order to advance the development of multilingual speech translation systems in the Asian region. Thus, the final speech translation system is expected to include not only Japanese, English, Chinese, and Indonesian, but other Asian languages as well. Future work in this field involves objective and subjective evaluation of the overall system in comparison with other language pairs speech-to-speech translation systems.

References

[1] W.J. Hutchins, "Machine translation: a brief history," in *Concise history of the language sciences: from the Sumerians to the cognitivists*, E.F.K.Koerner and R.E.Asher, Eds. 1995, pp. 431–445, Oxford: Pergamon Press.

[2] 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 2008*, Manchester, UK, 2008, pp. 21–24.

[3] T. Shimizu, Y. Ashikari, E. Sumita, J. Zhang, and S. Nakamura, "NICT/ATR Chinese-Japanese-English speech-to-speech translation system," *Tsinghua Science and Technology*, vol. 13, no. 4, pp. 540–544, 2008.

[4] S. Sakti, P. Hutagaol, A. Arman, and S. Nakamura, "Indonesian speech recognition for hearing and speaking impaired people," in *Proc. ICSLP*, Jeju, Korea, 2004, pp. 1037–1040.

[5] S. Sakti, K. Markov, and S.Nakamura, "Rapid development of initial indonesian phoneme-based speech recognition using cross-language approach," in *Proc. Oriental COCOSDA*, Jakarta, Indonesia, 2005, pp. 38–43.

[6] 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.

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

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

[9] 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.

[10] 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.

[11] S. Banerjee and A. Lavie, "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*. 2005, pp. 65–72, Aan Arbor, Michigan, USA.

[12] E. Sumita, T. Shimizu, and S. Nakamura, "NICT-ATR speech-to-speech translation system," in *Proc. ACL*, Prague, Czech Republic, 2007, pp. 25–28.

[13] K. Tokuda, T. Kobayashi, and S. Imai, "Adaptive cepstral analysis of speech," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 6, pp. 481–489, 1995.

[14] 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.