Toward Translating Indonesian Spoken Utterances to/from Other Languages

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

This paper outlines the National Institute of Information and Communications Technology / Advanced Telecommunications Research Institute International (NICT/ATR) research activities in developing a spoken language translation system, specially for translating Indonesian spoken utterances into/from Japanese or English. Since the NICT/ATR Japanese-English speech translation system is an established one and has been widely known for many years, our focus here is only on the additional components that are related to the Indonesian spoken language technology. This includes the development of an Indonesian large vocabulary continuous speech recognizer, Indonesian-Japanese and Indonesian-English machine translators, and an Indonesian speech synthesizer. Each of these component technologies was developed by using corpus-based speech and language processing approaches. Currently, all these components have been successfully incorporated into the mobile terminal of the NICT/ATR multilingual speech translation system.

1. Introduction

With the increasing trends of globalization, international tourism and international business, the issue of communication between people who do not share a common language becomes important. A common dream is to realize a speech translation technology that is able to break down the language barrier and enable instant cross-lingual communication. Many researchers have been working in the areas of speech recognition, machine translation and speech synthesis for nearly five decades. Owing to the advancements made in the field of spoken language processing technology, the dream of realizing a multilingual speech translation system has become more feasible. Recently, NICT/ATR successfully launched the first commercial speech translation service between the Japanese and English languages for real environments [1].

Indonesia, as the fourth most populous nation in the world, continues to be plagued by inadequate speech-enabling technology and research. Most language-related researchers in Indonesia play a limited role: they are only active in the fields of speech synthesis technologies and natural language processing. One of the main problems of developing an Indonesian speech recognition system is the lack of an Indonesian speech corpus. There are difficulties in developing such corpora because the Indonesian language is actually a "language of national unity" formed from the hundreds of languages spoken in the Indonesian archipelago. With more than 230 million people, only 7% of the population speak the Indonesian language as a mother tongue, while the great majority of people speak it as a second language with varying degrees of proficiency. The process of collating all of the possible speaking styles, with the innumerable mother tongues and dialects recognized in Indonesia, is still too difficult to be accomplished.

This paper describes the recent progress in the research on technological applications centered on the Indonesian spoken language, including the development of the Indonesian large-vocabulary corpora, the Indonesian large vocabulary continuous speech recognition (LVCSR) system, the Indonesian-Japanese and Indonesian-English machine translation systems, as well as the development of the Indonesian speech synthesis system. First, we briefly describe the Indonesian language corpora in section 2. Then, we describe the work and experiments related to the Indonesian LVCSR system (section 3), the machine translation system (section 4), and the development of the Indonesian speech synthesis system (section 5). The integration of these component technologies in a

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

hand-held speech translation system is described in section 6. Finally, we draw our conclusions in section 7.

2. Indonesian Language Corpora

Three types of Indonesian language corpora, available in both text and speech forms, were used here: these pertained to travel expressions, daily news, and telephone tasks. The first corpora was part of the ATR basic travel expression corpus (BTEC), while the other two corpora were 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 [2]. These corpora are described below.

2.1. Text Corpora

a. Travel expression task

The ATR basic travel expression corpus (BTEC) has served as the primary source for developing broad-coverage speech translation systems [3]. The sentences were collected by bilingual travel experts from Japanese/English sentence pairs in travel domain "phrasebooks." The ATR-BTEC has also been translated into 18 different languages, including French, German, Italian, Chinese, Korean, and Indonesian. Each language comprises 160,000 sentences (with about 20,000 unique words) of training, and a test set of 510 sentences with 16 references per sentence.

b. Daily news task

A raw text source for the daily news task has already been generated by an Indonesian student [4]. The source was compiled from "KOMPAS" and "TEMPO," which are currently the biggest and most widely read Indonesian newspaper and magazine, respectively. This source consists of more than 3160 articles, with around 600,000 sentences. R&D TELKOM further processed the raw text source to generate a clean text corpus.

c. Telephone application task

A total of 2500 sentences related to the telephone application domain were generated by R&D TELKOM. These were derived from the dialog that is commonly employed in the telephone services, including tele-home security, billing information services, reservation services, status tracking of e-Government services and hearing impaired telecommunication services (HITS).

2.2. Speech Corpora

2.2.1. Multi speakers

a. Travel expression task

From the test set of the BTEC text data described above, we chose 510 sentences single-reference test set by selecting the first reference from the 16 references. The speech recordings were done by the ATR in Jakarta, Indonesia. The Agency for Assessment and Application Technology (BPPT), Indonesia, also helped to investigate the preliminary recording. The clean speech recording was conducted in a sound proof room, at a 48 kHz sampling rate with 16-bit resolution. The sampling rate was later downsampled to 16 kHz for our experiments. We only selected speakers who spoke standard Indonesian (without an accent). There were 42 speakers (20 males, 22 females) and each speaker uttered the same set of 510 BTEC sentences, resulting in a total of 21,420 utterances (23.4 hours of speech).

b. Daily news task

From the text data of the news task described above, we selected phonetically-balanced sentences by using the greedy search algorithm [5]; this produced a total of 3168 sentences. Then, clean and telephone speech were recorded, simultaneously, at sampling frequencies of 16 and 8 KHz, respectively, by R&D TELKOM in Bandung, Java Island, Indonesia. There were a total of 400 speakers (200 males and 200 females). Four main accents were covered: Batak, Java, Sunda, and standard Indonesian (without accent). Each speaker uttered 110 sentences, resulting in a total of 44,000 speech utterances, which amounted to around 43.35 hours of speech.

c. Telephone application task

Using the same recording set-up as the one used for the news task corpus, the speech utterances for 2500 sentences pertaining to telephone application tasks were recorded by R&D TELKOM in Bandung, Indonesia; the total number of speakers and their distribution in terms of age and accents were also identical. Each speaker uttered 100 sentences, resulting in a total of 40,000 utterances (36.15 hours of speech).

Table 1. Recognition accuracy rates of Japanese, English and Indonesian LVCSR on BTEC test set

Language	# Phoneme	# Accent	# Speakers (M,F)	# Utterances	# Hours	Accuracy
Japanese	26	No Accent	4200 (1600, 2600)	172, 674	270.9	94.87%
English	44	3 (US, BRT, AUS)	532 (266, 266)	207,724	202.0	92.29%
Indonesian	33	4 (JV, SN, BT, ST)	400 (200, 200)	84,000	79.5	92.47%

2.2.2. Single speaker

We selected another set of phonetically-balanced sentences from both travel expressions and daily news texts by using, again, the greedy search algorithm. There were 2,012 sentences in total that were uttered by a female Indonesian speaker who spoke standard Indonesian, without an accent. The speech was conducted in a sound proof room, at a sampling rate of 48 kHz, with 16-bit resolution. The sampling rate was later downsampled to 16 kHz for our experiments.

3. Indonesian Speech Recognition

The Indonesian LVCSR system was trained by using the multi-speaker clean speech data on the daily news and telephone application tasks. The experiments were conducted by using the following feature extraction parameters: a sampling frequency of 16 kHz, the frame length of a 20 ms Hamming window, a frame shift of 10 ms, and 25 dimensional MFCC features (12-order MFCC, Δ MFCC, and Δ log power).

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 [6].

Using the resulting segmented utterance, we trained the acoustic model. Three states were used as the initial hidden Markov model (HMM) for each phoneme. A shared state HMnet topology was then obtained by using a successive state splitting (SSS) training algorithm based on the minimum description length (MDL) optimization criterion [7]. The resulting context-dependent triphone had 1,277 states in total, with 15 Gaussian mixture components per state.

Word bigram and trigram language models were trained by using the 160K sentences of the BTEC training 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 dictionary, which is owned by R&D TELKOM, was derived from the daily news and telephone application text corpora. It consists of about 40,000 words in total, including 30,000 original Indonesian words, 8000 names of people and places, and 2000 foreign words. The pronunciations of these words were manually developed by Indonesian linguists. Based on these pronunciations, we included additional words derived from the BTEC sentences.

The Indonesian speech recognition system achieved a performance of 92.47% word accuracy, at RTF=0.97 on the BTEC speech data. Table 1 shows a comparison between the performances of the Indonesian speech recognizer and speech recognizers pertaining to other languages.

4. Statistical Machine Translation

Phrase-based statistical machine translation systems for Indonesian-Japanese and Indonesian-English translations were trained by using 160K sentence pairs of BTEC text data. Monolingual features like the language model probability were trained on 160K monolingual text corpora of the target languages. For decoding, a multi-stack phrase-based SMT decoder called CleopATRa [8] was used.

The qualities of the Indonesian-Japanese and Indonesian-English SMT engines were evaluated by using sets of 510 sentences of test data, with 16 references per sentence. Table 2 shows the bilingual evaluation under-study (BLEU) [9] scores for both Indonesian-Japanese and Indonesian-English SMT engines in comparison with the Japanese-English SMT engine. More details on the NICT/ATR speech translation system can be found in [1, 10, 11].

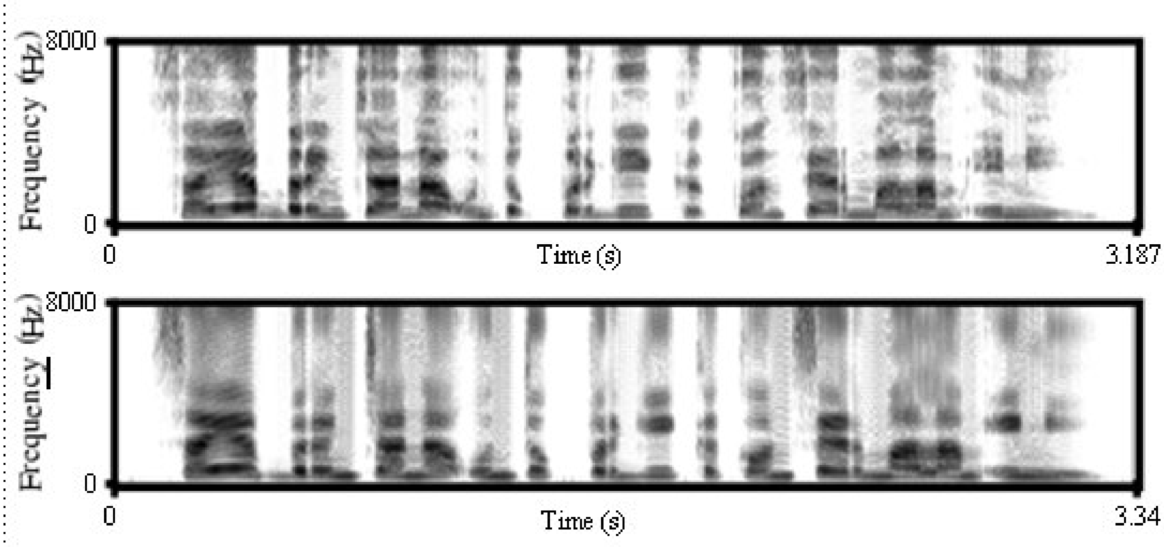

Figure 1. Spectrograms of both natural speech (top) and synthesized speech (bottom) for an utterance "Saya berencana untuk pergi ke konser malam ini" (meaning "I plan on going to the concert this evening").

Table 2. Automatic evaluation of SMT engines

Language pair	BLEU (%)
Indonesian-to-Japanese	57.24
Japanese-to-Indonesian	40.59
Indonesian-to-English	59.69
English-to-Indonesian	48.35
Japanese-to-English	61.56
English-to-Japanese	68.53

5. Indonesian Speech Synthesis

The Indonesian HMM-based speech synthesis system was trained using two hours of single-speaker phonetically-balanced speech data. 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 F0 and its dynamic parameters (delta and acceleration). The spectral feature vector consisted of 25 mel-cepstral coefficients [12], including the zeroth coefficient, and their dynamic parameters (delta and acceleration).

The full contextual label was generated from a phonetic transcription by using text processing tools, and it comprised only the phoneme identity and its positional information with regard to the words and sentences that were taken into account. This study did not use syllables or stress information. Part of speech (POS) tagging, too, was not included at this stage.

Five-state left-to-right HMMs were used, where each HMM corresponds to a phoneme-sized speech unit. These context-dependent HMMs were trained using the full contextual labels, and the concatenated feature vectors of the extracted F0 and mel-cepstrum parameters. The mel-cepstrum feature vectors were modeled by a continuous probability distribution, while the F0 feature vectors were modeled by a multi-spaced probability distribution (including a discrete voiced/unvoiced symbol and one-dimensional continuous log F0 values). The state durations of each HMM were modeled by n-dimensional Gaussians, whose dimensions were equal to the number of states of the HMM. Here, the distributions of the excitation (pitch) parameter, spectral parameter, and the 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 the spectrum, pitch and duration models had 2,409 leaves, 4,245 leaves and 961 leaves, respectively.

Speech waveform was synthesized by using simple excitation and the MLSA (mel-log spectrum approximation) filter [12]. Figure 1 shows an example of the spectrogram comparisons of natural (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 set. It was observed that the system was able to synthesize speech that resembled the recorded speech in the database. The speaking rate of the synthesized version was also similar to that of the natural speech recording. Through informal listening tests, we found that the synthesized speech continues to present a characteristic buzziness, caused by the simple excitation model. However, by and large, the prosody is good and the speech sounds smooth and stable. In the future, we will improve the naturalness by utilizing a wider contextual factors such as syllables, stress, phrases and POS tags, as well as the incorporation of a prosodic break module into text processing front end.

6. Integration in a Handheld Speech Translation System

The Indonesian LVCSR system, the Indonesian-Japanese and Indonesian-English machine translation systems, as well as the Indonesian HMM-based speech synthesis system described above, have presently been integrated into the mobile terminal of the NICT/ATR multilingual speech translation system. It is designed for practical use as a translation assistance tool for travelers going abroad. Figure 2 shows the entire speech translation system, which consists of speech recognition modules, machine translation modules and speech synthesis modules. This system will translate utterance by utterance in a real-world environment. The input speech of the source language is recognized by using the speech recognizer. Then, the resulting text is translated into the target language by the machine translator. Finally, the synthesizer is used to produce the spoken target output.

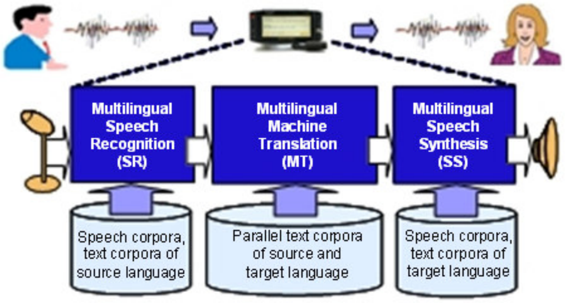
Multilingual Speech Recognition (SR)
Multilingual Machine Translation (MT)
Multilingual Speech Synthesis (SS)
Speech corpora, text corpora of source language
Parallel text corpora of source and target language
Speech corpora, text corpora of target language

Figure 2. An architecture of the NICT/ATR Indonesian-Japanese speech translation system, including speech recognition, machine translation and speech synthesis.

Part of this project was carried out in accordance with the Asia speech translation (A-STAR) consortium [13]. 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 the Japanese, English, and Indonesian languages, but also many other languages from other Asian countries. These fundamental technologies are expected to be applicable to the human-machine interfaces of various telecommunication devices and services connecting many Asian countries through a network. The improvements in borderless communication in the Asian region are expected to produce benefits in many fields, including tourism, business, education and social life.

7. Conclusion

We have presented an account of the development of the Indonesian spoken language technologies, including the Indonesian LVCSR system, the Indonesian-Japanese and Indonesian-English machine translation systems, as well as the Indonesian speech synthesis system. Each of these component technologies was developed by using a corpus-based approach. The experimental results show that Indonesian speech recognition in the travel domain was 92.47% accurate, with BLEU scores of 57.24% and 59.69% for Indonesian-Japanese and Indonesian-English machine translations, respectively. Moreover, the quality of the Indonesian synthesized speech was found to be smooth and perfectly intelligible. Finally, these component systems were successfully integrated into a hand-held NICT/ATR multi-lingual speech translation system.

References

[1] 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.

[2] 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.

[3] 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.

[4] F. Tala, A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia, Ph.D. thesis, The Information

and Language System (ILPS) Group, Informatics Institute, University of Amsterdam, Amsterdam, Netherland, 2003.

[5] J. Zhang and S.Nakamura, “An efficient algorithm to search for a minimum sentence set for collecting speech database,” in Proc. ICPhS, Barcelona, Spain, 2003, pp. 3145 -3148.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.