Development of Indonesian Large Vocabulary Continuous Speech Recognition System within A-STAR Project

Sakriani Sakti[1,2], **Eka Kelana**[3], **Hammam Riza**[4], **Shinsuke Sakai**[1,2]
Konstantin Markov[1,2], **Satoshi Nakamura**[1,2]
[1]National Institute of Information and Communications Technology, Japan
[2]ATR Spoken Language Communication Research Laboratories, Japan
[3]R&D Division, PT Telekomunikasi Indonesia, Indonesia
[4]Agency for the Assessment and Application of Technology, BPPT, Indonesia
{sakriani.sakti,shinsuke.sakai,konstantin.markov,satoshi.nakamura}@atr.jp,
eka_k@telkom.co.id, hammam@iptek.net.id

Abstract

The paper outlines the development of a large vocabulary continuous speech recognition (LVCSR) system for the Indonesian language within the Asian speech translation (A-STAR) project. An overview of the A-STAR project and Indonesian language characteristics will be briefly described. We then focus on a discussion of the development of Indonesian LVCSR, including data resources issues, acoustic modeling, language modeling, the lexicon, and accuracy of recognition. There are three types of Indonesian data resources: daily news, telephone application, and BTEC tasks, which are used in this project. They are available in both text and speech forms. The Indonesian speech recognition engine was trained using the clean speech of both daily news and telephone application tasks. The optimum performance achieved on the BTEC task was 92.47% word accuracy.

1 A-STAR Project Overview

The A-STAR project is an Asian consortium that is expected to advance the state-of-the-art in multilingual man-machine interfaces in the Asian region. This basic infrastructure will accelerate the development of large-scale spoken language corpora in Asia and also facilitate the development of related fundamental information communication technologies (ICT), such as multi-lingual speech translation, multi-lingual speech transcription, and multi-lingual information retrieval.

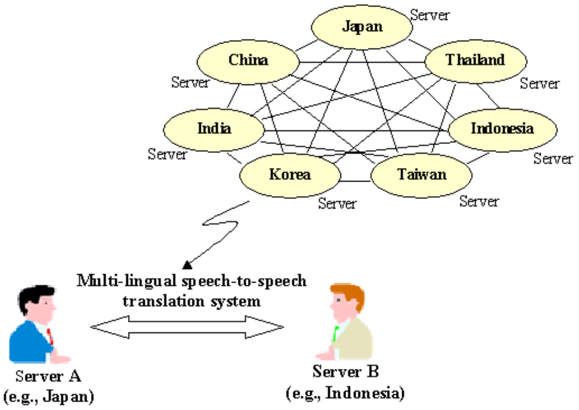

Figure 1: Outline of future speech-technology services connecting each area in the Asian region through network.

These fundamental technologies can be applied to the human-machine interfaces of various telecommunication devices and services connecting Asian countries through the network using standardized communication protocols as outlined in Fig. 1. They are expected to create digital opportunities, improve our digital capabilities, and eliminate the digital divide resulting from the differences in ICT levels in each area. The improvements to borderless communication in the Asian region are expected to result in many benefits in everyday life including tourism, business, education, and social security.

The project was coordinated together by the Advanced Telecommunication Research (ATR) and the

National Institute of Information and Communications Technology (NICT) Japan in cooperation with several research institutes in Asia, such as the National Laboratory of Pattern Recognition (NLPR) in China, the Electronics and Telecommunication Research Institute (ETRI) in Korea, the Agency for the 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 Taiwan University (NTU) in Taiwan. Partners are still being sought for other languages in Asia.

More details about the A-STAR project can be found in (Nakamura et al., 2007).

2 Indonesian Language Characteristic

The Indonesian language, or so-called Bahasa Indonesia, is a unified language formed from hundreds of languages spoken throughout the Indonesian archipelago. Compared to other languages, which have a high density of native speakers, Indonesian is spoken as a mother tongue by only 7% of the population, and more than 195 million people speak it as a second language with varying degrees of proficiency. There are approximately 300 ethnic groups living throughout 17,508 islands, speaking 365 native languages or no less than 669 dialects (Tan, 2004). At home, people speak their own language, such as Javanese, Sundanese or Balinese, even though almost everybody has a good understanding of Indonesian as they learn it in school.

Although the Indonesian language is infused with highly distinctive accents from different ethnic languages, there are many similarities in patterns across the archipelago. Modern Indonesian is derived from the literary of the Malay dialect. Thus, it is closely related to the Malay spoken in Malaysia, Singapore, Brunei, and some other areas.

Unlike the Chinese language, it is not a tonal language. Compared with European languages, Indonesian has a strikingly small use of gendered words. Plurals are often expressed by means of word repetition. It is also a member of the agglutinative language family, meaning that it has a complex range of prefixes and suffixes, which are attached to base words. Consequently, a word can become very long.

More details on Indonesian characteristics can be found in (Sakti et al., 2004).

3 Indonesian Phoneme Set

The Indonesian phoneme set is defined based on Indonesian grammar described in (Alwi et al., 2003). A full phoneme set contains 33 phoneme symbols in total, which consists of 10 vowels (including diphthongs), 22 consonants, and one silent symbol. The vowel articulation pattern of the Indonesian language, which indicates the first two resonances of the vocal tract, F1 (height) and F2 (backness), is shown in Fig. 2.

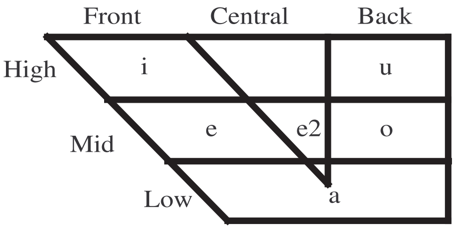

Figure 2: Articulatory pattern of Indonesian vowels.

It consists of vowels, i.e., /a/ (like “a” in “father”), /i/ (like “ee” in “screen”), /u/ (like “oo” in “soon”), /e/ (like “e” in “bed”), /e2/ (a schwa sound, like “e” in “learn”), /o/ (like “o” in “boss”), and four diphthongs, /ay/, /aw/, /oy/ and /ey/. The articulatory pattern for Indonesian consonants can be seen in Table 1.

4 Indonesian Data Resources

Three types of Indonesian data resources available in both text and speech forms were used here. The first two resources were developed or processed by the R&D Division of PT Telekomunikasi Indonesia (R&D TELKOM) in collaboration with ATR as continuation of the APT project (Sakti et al., 2004), while the third one was developed by ATR under the A-STAR project in collaboration with BPPT. They are described in the following.

Table 1: *Articulatory pattern of Indonesian consonants.*

	Bilabial	Labiodental	Dental/Alveolar	Palatal	Velar	Glotal
Plosives	p, b		t, d		k, g	
Affricates				c, j		
Fricatives		f	s, z	sy	kh	h
Nasal	m		n	ny	ng	
Trill			r			
Lateral			l			
Semivowel	w			y		

4.1 Text Data

The three text corpora are:

1. Daily News Task
   There is already a raw source of Indonesian text data, which has been generated by an Indonesian student (Tala, 2003). The source is a compilation from "KOMPAS" and "TEMPO", which are currently the largest and most widely read Indonesian newspaper and magazine. It consists of more than 3160 articles with about 600,000 sentences. R&D TELKOM then further processed them to generate a clean text corpus.

2. Telephone Application Task
   About 2500 sentences from the telephone application domain were also generated by R&D TELKOM, and were derived from some daily dialogs from telephone services, including tele-home security, billing information services, reservation services, status tracking of e-Government services, and also hearing impaired telecommunication services (HITSs).

3. BTEC Task
   The ATR basic travel expression corpus (BTEC) has served as the primary source for developing broad-coverage speech translation systems (Kikui et al., 2003). The sentences were collected by bilingual travel experts from Japanese/English sentence pairs in travel domain "phrasebooks". BTEC has also been translated into several languages including French, German, Italian, Chinese and Korean. Under the A-STAR project, there are also plans to collect synonymous sentences from the different languages of the Asian region. ATR has currently successfully collected an Indonesian version of BTEC tasks, which consists of 160,000 sentences (with about 20,000 unique words) of a training set and 510 sentences of a test set with 16 references per sentence. There are examples of BTEC English sentences and synonymous Indonesian sentences in Table 2.

Table 2: *Examples of English-Indonesian bilingual BTEC sentences.*

English	**Indonesian**
Good Evening	Selamat Malam
I like strong coffee	Saya suka kopi yang kental
Where is the boarding gate?	Di manakah pintu keberangkatan berada?
How much is this?	Harganya berapa?
Thank you	Terima kasih

4.2 Speech Data

The three speech corpora are:

1. Daily News Task
   From the text data of the news task described above, we selected phonetically-balanced sentences, then recorded the speech utterances. Details on the phonetically-balanced sentences, the recording set-up, speaker criteria, and speech utterances are described in what follows:

   - Phonetically-Balanced Sentences
     We selected phonetically-balanced sentences using the greedy search algorithm

(Zhang and S.Nakamura, 2003), resulting in 3168 sentences in total (see Table 3).

Table 3: *Number of phonetically-balanced sentences resulting from greedy search algorithm.*

Phone	# Units	# Sentences
Monophones	33	6
Left Biphones	809	240
Right Biphones	809	242
Triphones	9667	2978
Total		3168

- Recording Set-Up
  Speech recording was done by R&D TELKOM in Bandung, Java, Indonesia. It was conducted in parallel for both clean and telephone speech, recorded at respective sampling frequency of 16 kHz and 8 kHz. The system configuration is outlined in Fig. 3.

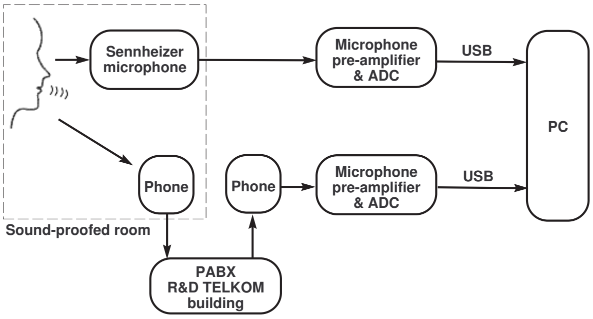

Figure 3: Recording set-up.

- Speaker Criteria
  The project will require a lot of time, money, and resources to collect all of the possible languages and dialects of the tribes recognized in Indonesia. In this case, R&D TELKOM only focused on the major ethnic accents in Bandung area where the actual telecommunication services will be implemented. Four main accents were selected, including: Batak, Javanese, Sundanese, and standard Indonesian (no accent) with appropriate distributions as outlined in Fig. 4. Both genders are evenly distributed and the speakers' ages are also distributed as outlined in Fig. 5. The largest percentage is those aged 20-35 years who are expected to use the telecommunication services more often.

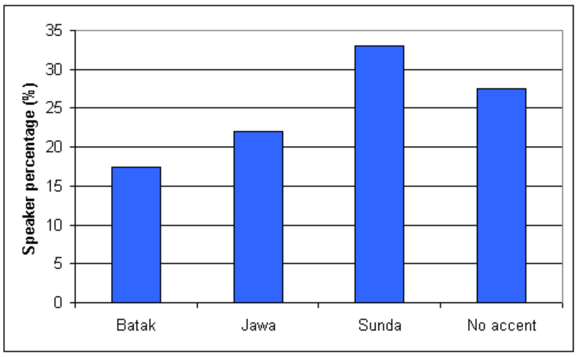

Figure 4: Accent distribution of 400 speakers in daily news and telephone application tasks.

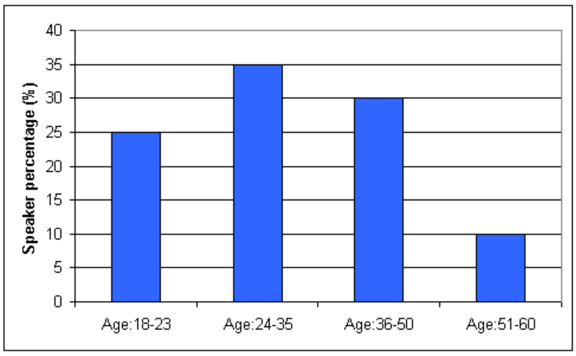

Figure 5: Age distribution of 400 speakers in daily news and telephone application tasks.

- Speech Utterances
  The total number of speakers was 400 (200 males and 200 females). Each speaker uttered 110 sentences resulting in a total of 44,000 speech utterances or about 43.35 hours of speech.

2. Telephone Application Task
   The utterances in speech of 2500 telephone application sentences were recorded by R&D TELKOM in Bandung, Indonesia using the same recording set-up as that for the news task

corpus. The total number of speakers, as well as appropriate distributions for age and accent, were also kept the same. Each speaker uttered 100 sentences resulting in a total of 40,000 utterances (36.15 hours of speech).

3. BTEC Task
   From the test set of the BTEC text data previously described, 510 sentences of one reference were selected and the recordings of speech were then done by ATR in Jakarta, Indonesia. BPPT helped to evaluate the preliminary recordings. For this first version, we only selected speakers who spoke standard Indonesian (no accent). There were 42 speakers (20 males and 22 females) and each speaker uttered the same 510 BTEC sentences, resulting in a total of 21,420 utterances (23.4 hours of speech).

5 Indonesian Speech Recognizer

The Indonesian LVCSR system was developed using the ATR speech recognition engine. The clean speech of both daily news and telephone application tasks were used as the training data, while the BTEC task was used as an evaluation test set. More details on the parameter set-up, acoustic modeling, language modeling, pronunciation dictionary and recognition accuracy will be described in the following.

5.1 Parameter Set-up

The experiments were conducted using feature extraction parameters, which were a sampling frequency of 16 kHz, a 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).

5.2 Segmentation Utterances

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. In this case, we used an available Indonesian phoneme-based acoustic model developed using the English-Indonesian cross language approach (Sakti et al., 2005).

5.3 Acoustic Modeling

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 (Jitsuhiro et al., 2004). Various MDL parameters were evaluated, resulting in context-dependent triphone systems having different version of total states. i.e., 1,277 states, 1,944 states and 2,928 states. All triphone HMnets were also generated with three different versions of Gaussian mixture components per state, i.e., 5, 10, and 15 mixtures.

5.4 Language Modeling

Word bigram and trigram language models were trained using the 160,000 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. This high perplexity could be due to agglutinative words in the Indonesian language.

5.5 Pronunciation Dictionary

About 40,000 words from an Indonesian pronunciation dictionary were manually developed by Indonesian linguists and this was owned by R&D TELKOM. This was derived from the daily news and telephone application text corpora, which consisted of 30,000 original Indonesian words plus 8,000 person and place names and also 2,000 of foreign words. Based on these pronunciations, we then included additional words derived from the BTEC sentences.

5.6 Recognition Accuracy

The performance of the Indonesian speech recognizer with different versions of total states and Gaussian mixture components per state is graphically depicted in Fig. 6. On average, they achieved 92.22% word accuracy. The optimum performance was 92.47% word accuracy at RTF=0.97 (XEON 3.2 GHz), which was obtained by the model with 1.277 total states and 15 Gaussian mixture components per state.

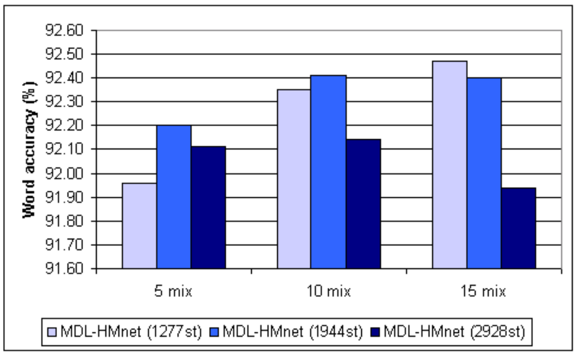

Figure 6: Recognition accuracy of Indonesian LVCSR on BTEC test set.

6 Conclusion

We have presented the results obtained from the preliminary stages of an Indonesian LVCSR system. The optimum performance achieved was 92.47% word accuracy at RTF=0.97. A future development will be to implement it on a real speech-to-speech translation system using computer terminals (tablet PCs). To further refine the system, speaker adaptation as well as environmental or noise adaptation needs to be done in the near future.

References

H. Alwi, S. Dardjowidjojo, H. Lapoliwa, and A.M. Moeliono. 2003. *Tata Bahasa Baku Bahasa Indonesia (Indonesian Grammar)*. Balai Pustaka, Jakarta, Indonesia.

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.

G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto. 2003. Creating corpora for speech-to-speech translation. In *Proc. EUROSPEECH*, pages 381–384, Geneva, Switzerland.

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*, page to appear, Yamanashi, Japan.

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.

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

F. Tala. 2003. *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.

J. Tan. 2004. Bahasa indonesia: Between faqs and facts. http://www.indotransnet.com/article1.html.

J. Zhang and S.Nakamura. 2003. An efficient algorithm to search for a minimum sentence set for collecting speech database. In *Proc. ICPhS*, pages 3145–3148, Barcelona, Spain.