Large Vocabulary ASR for Indonesian Language in the A-STAR Project *

○ Sakriani Sakti(NICT/ATR-SLC), Eka Kelana (R&D TELKOM),
Hammam Riza (BPPT), Satoshi Nakamura (NICT/ATR-SLC)

1 Introduction

In this paper, we present the development of an Indonesian large vocabulary continuous speech recognition (LVCSR) system within the Asia speech translation (A-STAR) project. An overview of the A-STAR Project and Indonesian language characteristics will be briefly described. Then, we focus on the discussion of Indonesian LVCSR development, including data resources issues, acoustic modeling, language modeling, lexicon and recognition performance.

2 The A-STAR Project

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

The project is coordinated by the Advanced Telecommunication Research (ATR) 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, and seek partners for other languages in Asia.

More details about A-STAR project can be found in [1].

3 Indonesian Language Characteristic

The Indonesian language, so-called Bahasa Indonesia, is a unity language formed from hundreds of languages spoken in 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 profiency. It is originally derived from the literary of the Malay dialect. Thus, it is closely related to 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 familly, meaning that it has a complex range of prefixes and suffixes which are attached to base words. So a word can become very long.

More details of Indonesian characteristics can be found in [2].

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 R&D Division of PT Telekomunikasi Indonesia (R&D TELKOM) in collaboration with ATR as continuation of APT project [2], while the last one was developed by ATR under A-STAR project in collaboration with BPPT. They are described in the following.

4.1 Text Data

1. Daily News Task
   There is already a raw source of Indonesian text data in the news domain which is generated by [3]. The source was a compilation from "KOMPAS" and "TEMPO", the current biggest and widely used of the Indonesian newspapers and magazine. It consists of more than 3160 articles with about 600,000 sentences. R&D TELKOM was then further processed to generate a clean text corpus.
2. Telephone Application Task
   About 2500 sentences of telephone application domain was also generated by R&D TELKOM, and were derived from some necessary dialogs of telephone services, including tele-home security, billing information service, reservation service, status tracking of e-Government service and also hearing impaired telecomuniccation services (HITS).
3. BTEC Task
   In the A-STAR project, ATR has collected an Indonesian version of the basic travel expression corpus (BTEC) [4]. It consists of about 160,000 sentences with about 20,000 unique words.

4.2 Speech Data

- Daily News Task
  From the text data of the news task described above, we selected phonetically-balanced sentences using a greedy search algorithm [5]. Then, speech recording was done by R&D TELKOM in Bandung, Indonesia. It was conducted in parallel for both clean and telephone speech, recorded in 16kHz and 8KHz sampling frequency, respectively. The total number of speakers is 400 (200 males, 200 females) with approriate distribution of age and four main accents: Batak, Java, Sunda, and standard Indonesian (no accent) as oulined in Fig. 1 and 2. Each speaker utterred 110 sentences resulting in a total of 44,000 utterances.
- Telephone Application Task
  Similar to the news task corpus, the speech utterances of 2500 telephone application sentences were recorded by R&D TELKOM in Bandung, Indonesia, using the same recording set-up. Each speaker utterred 100 sentences resulting in a total of 40,000 utterances.
- BTEC Task
  In the A-STAR project, ATR has also sucessfully collected the speech corpus of BTEC task. The collection was done in Jakarta, Indonesia, where BPPT gave help to investigate the preliminary

[*] A - STAR プロジェクトにおけるインドネシア語の大語彙音声認識,
サクテイサクリアニ (NICT/ATR-SLC), エカケラナ (R&D TELKOM),
ハマムリザ (BPPT), 中村 哲 (NICT/ATR-SLC)

recording. It consists of a total of 21,420 utterances spoken by 42 speakers (20 males, 22 females) where each speaker uttered 510 BTEC sentences.

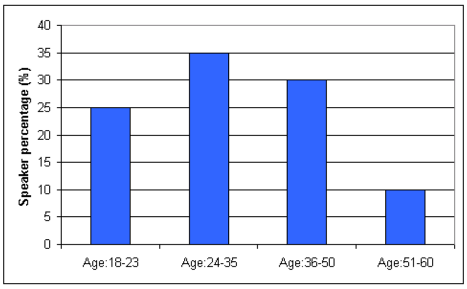

Fig. 1 Age distribution of 400 speakers of the daily news and telephone application task.

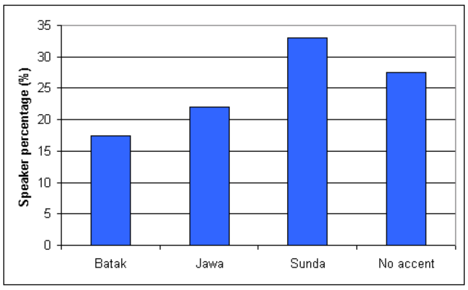

Fig. 2 Accent distribution of 400 speakers of the daily news and telephone application task.

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 speech corpus were used as a training data, while the BTEC speech corpus was used as an evaluation test set. More details about parameter set-up, acoustic modeling, language modeling, pronunciation dictionary and recognition accuracy are described in the following.

- Parameter Set-up
  The experments were conducted using feature extraction parameters that 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). The Indonesian phoneme set used here contains of 10 vowels (including diphthongs), 22 consonants, and 1 silence symbol. The articulation pattern of those phonemes can be found in [6].

- Acoustic Modeling
  A segmented utterance is produced by forced alignment using an available Indonesian phoneme-based acoustic model developed using the Engligh-Indonesian cross language approach [6]. 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 [7]. Various MDL parameters were investigated, 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.

- Language Modeling
  Word bigram and trigram language models were trained using the 160k BTEC text corpus, yielding a trigram perplexity of 67.0 and an out-of vocabulary (OOV) rate of 0.78% on the 510 BTEC test set. The high perplexity might be due to agglutination words in the Indonesian language.

- Pronunciation Dictionary
  About 40k words of Indonesian pronunciation dictionary was manually developed by Indonesian linguists and it is owned by R&D TELKOM. This was derived from the daily news and telephone application text corpus, which consists of 30k of original Indonesian words plus 8k of person and place names and also 2k of foreign words. Based on those pronunications, we then include additional words derived from 160k BTEC text corpus.

- Recognition Accuracy
  The performance of the Indonesian speech recognizer with different versions of total states and Gaussian mixture components per state is depicted in Fig. 3. 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.

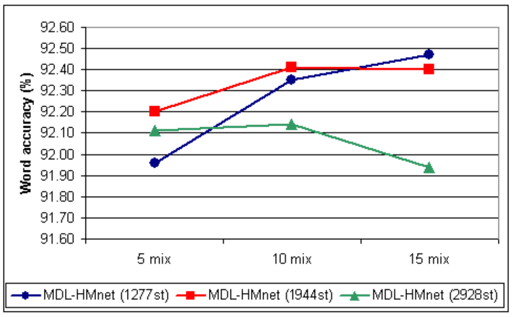

Fig. 3 Recognition accuracy of Indonesian LVCSR on the BTEC test set.

6 Conclusion

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

参考文献

[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, p. to appear.

[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] 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.

[4] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, "Creating corpora for speech-to-speech translation," in *Proc. EUROSPEECH*, Geneva, Switzerland, 2003, pp. 381–384.

[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.