Recent Progress in Developing Indonesian Large-Vocabulary Corpora and LVCSR System

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

Speech-related research in Indonesia has focused mainly on speech synthesizer technologies and natural language processing. More recently, work on collecting Indonesian speech corpora and developing speech recognition systems was initiated. However, most of these systems only recognized a very limited vocabulary. This paper outlines recent progress in developing Indonesian large-vocabulary corpora and a large vocabulary continuous speech recognition (LVCSR) system. Research on the Indonesian LVCSR has been carried out under the A-STAR (Asian Speech Translation Advanced Research) consortium. Three types of Indonesian large vocabulary data sources were used: daily news, telephone application and basic travel expression (BTEC) tasks, which are available in both text and speech forms. The Indonesian speech recognition engine was trained using clean speech for both the daily news and telephone application tasks, and the performance was evaluated using the BTEC task.

1. Introduction

With more than 230 million people, Indonesia is the fourth most populous nation in the world. After considering population, variability, distribution, religious circumstances and linguistic aspects, Indonesian/Malay was ranked ninth for inclusion in the Global-Phone speech database [1]. It also ranks highly in the speech science community [2]. These results are seemingly at odds with the fact that Indonesian language still suffers from inadequate speech-enabling technology and research.

One of the main problems is the lack of Indonesian speech corpora. 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. Only 7% of the population speak it as a mother tongue, while the great majority of people speak it as a second language with varying degrees of proficiency. It is still considered too difficult to attempt to collect all of the possible languages and dialects of the various ethnic groups recognized in Indonesia.

Some researchers proposed producing an Indonesian speech recognition system using cross-lingual pronunciation modeling based on other resource languages. However, it was observed that this would result in poor performance [3]. In 2004, an Indonesian speech-enabling technology project for a telecommunication system was initiated by the R&D Division of PT Telekomunikasi Indonesia (R&D TELKOM) [4]. This project successfully collected a small vocabulary of the Indonesian speech corpora for digit and simple dialog tasks, which covered several of the major ethnic dialects spoken in Indonesia. A word-based Indonesian speech recognition system was also developed. Finally, in 2006, research [5] began on developing a large-vocabulary Indonesian corpora and speech recognition system. However, only 20 speakers were used in collecting the corpora and each speaker uttered only 328 sentences (14.5 hours of speech in total), which is still too small for an optimum Indonesian LVCSR system.

This paper describes recent progress in research on the Indonesian language, which has resulted in the successful development of an Indonesian LVCSR system trained using about 80 hours of Indonesian large-vocabulary speech corpora. We first briefly describe the A-STAR project background in Section 2 and the overview of Indonesian phoneme set in Section 3. We then describe works and ex-

periments relating to the Indonesian large-vocabulary corpora (Section 4), initial segmentation issues (Section 5), and the development of the Indonesian LVCSR system (Section 6). Finally, conclusions are drawn in Section 7.

2 Project Background

The work was carried out under the A-STAR (Asian Speech Translation Advanced Research) consortium. This is a consortium of Asian speech and natural language researchers, 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. Current A-STAR members include the National Laboratory of Pattern Recognition (NLPR) in China, the Electronics and Telecommunication Research Institute (ETRI) in Korea, the Agency for 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, and the National Taiwan University (NTU) in Taiwan. Partners are still being sought for other languages in Asia.

The goal of the project is to advance the development of multi-lingual man-machine interfaces in the Asia region. This basic infrastructure will accelerate the development of large-scale spoken language corpora in Asia and also facilitate the development of related fundamental components of information communication technologies (ICT), especially multi-lingual speech translation systems.

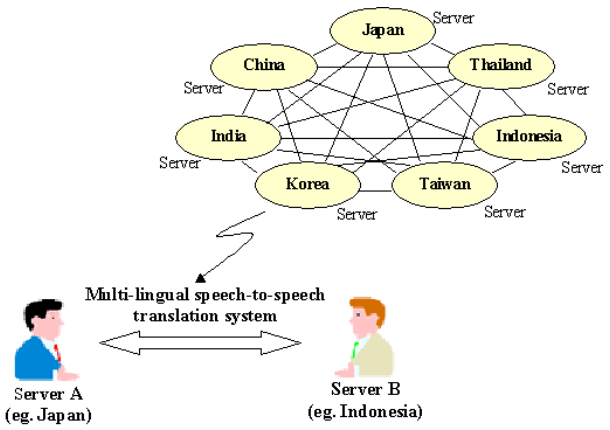

Figure 1. Outline of future speech-technology services connecting each area in the Asian region through network.

These fundamental 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 shown in Fig. 1. Improvements in borderless communication in the Asia region are expected to have benefits in many areas including tourism, business, education, and social security.

More details about the A-STAR consortium are available elsewhere [6].

3 Indonesian Phoneme Set

The Indonesian phoneme set is defined on the basis of an Indonesian grammar text [7]. A full phoneme set contains a total of 33 phoneme symbols, which consist of 10 vowels (including diphthongs), 22 consonants, and 1 silence 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.

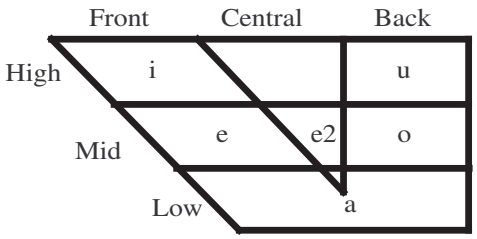

Figure 2. Articulatory pattern of Indonesian vowels.

The pattern consists of the vowels: /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 of Indonesian consonants is given in Table 1.

4 Indonesian Large Vocabulary Corpora

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 TELKOM in collaboration with ATR as a continuation of the APT (Asia Pacific Telecommunity) project [4], while the last one was developed by ATR under the A-STAR project in collaboration with BPPT. They are described below.

Table 1. Articulatory pattern of Indonesian consonants.

	Bilabial	Labiodental	Dental/Alveolar	Palatal	Velar	Glottal
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 Daily News Task

Text Corpora

A raw text source for the daily news task has already been generated by an Indonesian student [8]. The source was compiled from "KOMPAS" and "TEMPO", currently the biggest and most widely used of Indonesian newspapers and magazines. It consists of more than 3160 articles with about 600,000 sentences. R&D TELKOM further processed the raw text source to generate a clean text corpus.

Speech Corpora

From the text data of the news task described above, we selected phonetically-balanced sentences using the greedy search algorithm [9], which produced a total of 3168 sentences. Then, speech recording was conducted in parallel for both clean and telephone speech, at sampling frequencies of 16 and 8 KHz, respectively, by R&D TELKOM in Bandung, Java Island, Indonesia. The system configuration used is presented in Fig. 3.

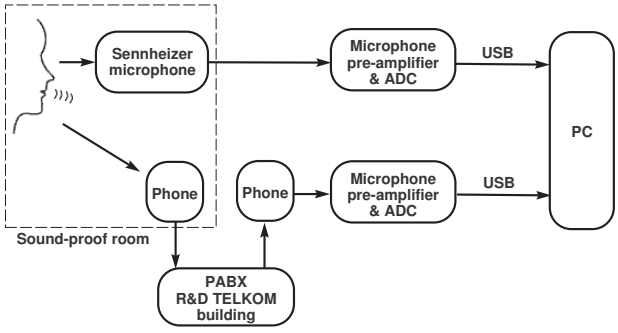

Figure 3. Recording set-up.

Collecting all the possible languages and dialects of all the ethnic groups recognized in Indonesia, would require a great deal of time, money, and resources. In this case, R&D TELKOM focused only on the ethnic accents most commonly found in the Bandung area where the actual telecommunication service will be implemented. Four main accents were selected: Batak, Java, Sunda, and standard Indonesian (no accent), with appropriate distribution as outlined in Fig. 4a. To obtain a variety of dialect accents, we asked each speaker to speak naturally without any pronunciation restriction. Both genders were distributed evenly and the speakers' ages were also distributed as outlined in Fig. 4b. The biggest percentage of speakers was aged 20-35 years with a Sunda accent. This group represented the people who are expected to be the biggest users of the telecommunication service.

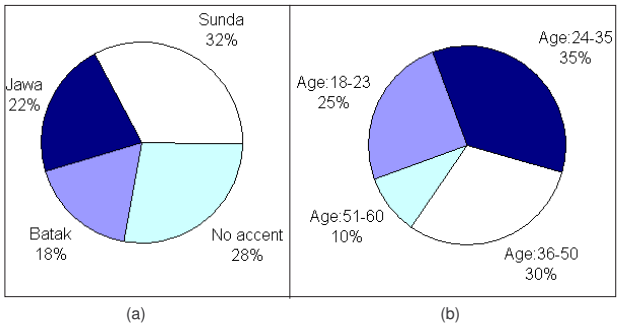

Figure 4. (a) Accent distribution of 400 speakers of daily news and telephone application tasks, and (b) age distribution of 400 speakers of daily news and telephone application tasks.

There was a total of 400 speakers (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.

4.2 Telephone Application Task

Text Corpora

With total of 2500 sentences from the telephone application domain were generated by R&D TELKOM. They were derived from some of the necessary dialogs used in telephone services, including tele-home security, billing information services, reservation services, status tracking of e-Government services and hearing impaired telecommunication services (HITS).

Speech Corpora

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

4.3 Basic Travel Expression Task

Text Corpora

The ATR basic travel expression corpus (BTEC) has served as the primary source for developing broad-coverage speech translation systems [10]. 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 Asia region. Currently, ATR has successfully collected an Indonesian version of the BTEC task, which consists of 160,000 sentences (with about 20,000 unique words) of training set and 510 sentences of test set with 16 references per sentence.

Speech Corpora

From the test set of BTEC text data described above, 510 sentences of one reference were selected and the recordings of clean speech was then done by ATR in Jakarta, Indonesia. BPPT helped to investigate the preliminary recording. For this first version, we selected only speakers who spoke standard Indonesian (no accent). 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).

5 Segmentation Issues

In developing speech recognition systems, utterances segmented according to labels are usually used as a starting point for training speech models. Automatic segmentation is generally used because it is efficient and faster. It is basically produced by forced alignment given the transcription. In this case, we first needed to develop an initial phoneme-based acoustic model.

The problem is that the available small-vocabulary Indonesian speech corpora [4] covers only 70% of the full Indonesian phoneme set. To obtain a proper Indonesian acoustic model with a full phoneme set, the missing phoneme models were inserted from the seed model developed using the English-Indonesian cross-language approach illustrated in Fig. 5.

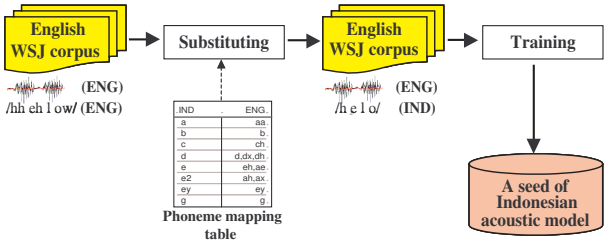

Figure 5. Seed acoustic model developed using English-Indonesian cross language approach.

First, we used English speech data comprising the Wall Street Journal (WSJ0 and WSJ1) speech corpus [11], substituted the English phoneme label alignments with Indonesian phoneme labels using the English-to-Indonesian phoneme mapping shown in Table 2, then trained the model as a seed of an Indonesian acoustic model.

Table 2. English-to-Indonesian phoneme mapping.

IND	ENG	IND	ENG	IND	ENG
a	aa	h	hh	oy	oy
ay	ay	i	ih,iy,ix	p	p
aw	aw	j	jh	r	r
b	b	k	k	s	s
c	ch	kh	k+h	sy	sh
d	d,dx,dh	l	l	t	t,th
e	eh,ae	m	m	u	uh,uw
e2	ah,ax	n	n	w	w
ey	ey	ng	ng	y	y
f	f,v	ny	n+y	z	z,zh
g	g	o	ow,ao	sil	sil

More details about the English-Indonesian cross-language approach can be found elsewhere [12].

6 Indonesian LVCSR System

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. The parameter set-up, acoustic modeling, language modeling, pronunciation dictionary and recognition accuracy are described more fully below.

6.1 Parameter Set-up

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).

6.2 Acoustic Modeling

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 (MDL) optimization criterion [13]. Various MDL parameters were investigated, resulting in context-dependent triphone systems that had different numbers of total states. i.e., 1,277, 1,944, 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.

6.3 Language Modeling

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

6.4 Pronunciation Dictionary

The dictionary, which is owned 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 people and place names and 2000 foreign words. All pronunciation of these words were manually developed by Indonesian linguists. Based on these pronunciations, we then included additional words derived from the BTEC sentences.

6.5 Recognition Accuracy

The performance of the Indonesian speech recognizer using different versions of total states and Gaussian mixture components per state is depicted in Fig. 6. On average, it achieved 92.22% word accuracy. The optimum performance was 92.47% word accuracy at RTF=0.97 (XEON 3.2 GHz), which was produced by the model with 1,277 total states and 15 Gaussian mixture components per state.

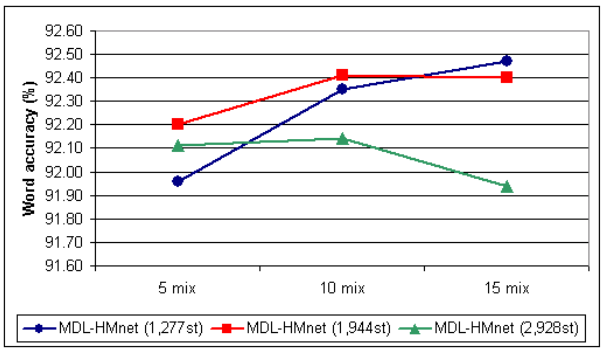

Figure 6. Recognition accuracy of Indonesian LVCSR on BTEC test set.

As a comparison, Table 3 shows the performance on BTEC test set of LVCSR system from several other languages, including Japanese, English and Chinese.

7 Conclusion

We have presented the current state of Indonesian large vocabulary corpora as well as the development of Indonesian LVCSR within the A-STAR project. An optimum performance of 92.47% word accuracy at RTF=0.97 was achieved. The next state of the work will include implementing the system on a real speech-to-speech translation system using computer terminals (tablet PCs). In the near future, the system will be further refined by incorporating speaker adaptation as well as environmental or noise adaptation.

8 Acknowledgement

A part of this A-STAR consortium activities has been supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT), the Asia Pacific Economic Cooperation Telecommunication and Information (APEC-TEL) Working Group, and the Asia Pacific Telecommunity Standardization Program (APT-ASTAP).

Table 3. Comparing recognition accuracy rates of different languages 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%
Chinese	85	4 (BJ, SH, CT, TW)	536 (268, 268)	207,257	249.2	90.65%
Indonesian	33	4 (JV, SN, BT, ST)	400 (200, 200)	84,000	79.5	92.47%

References

[1] T. Schultz, M. Westphal, and A. Waibel, "The global phone project: Multilingual lvcsr with janus-3," in *Proc. SQEL Workshop*, Pizen, Czech, 1997, pp. 20–27.

[2] E. Wong, T. Martin, T. Svendsen, and S. Sridharan, "Multilingual phone clustering for recognition of spontaneous indonesian speech utilising pronunciation modelling techniques," in *Proc. EUROSPEECH*, Geneva, Switzerland, 2003, pp. 3133–3136.

[3] T. Martin, T. Svendsen, and S. Sridharan, "Cross-lingual pronunciation modelling for indonesian speech recognition," in *Proc. EUROSPEECH*, Geneva, Switzerland, 2003, pp. 3125–3128.

[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] D.P. Lestari, K. Iwano, and S. Furui, "Development of an indonesian large vocabulary continuous speech recognition system," in *Proc. ASJ Autumn Meeting*, Kanazawa, Japan, 2006, pp. 41–42.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] D.B. Paul and J.M. Baker, "The design for the Wall Street journal based CSR corpus," in *Proc. DARPA Workshop*, Pacific Groove, California, USA, 1992.

[12] 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.

[13] 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.