RECENT PROGRESS IN DEVELOPING GRAPHEME-BASED SPEECH RECOGNITION FOR INDONESIAN ETHNIC LANGUAGES: JAVANESE, SUNDANESE, BALINESE AND BATAKS

Sakriani Sakti, Satoshi Nakamura

Augmented Human Communication Laboratory,
Graduate School of Information Science,
Nara Institute of Science and Technology, Japan
{ssakti,s-nakamura}@is.naist.jp

ABSTRACT

With the advent of globalization, multilingualism in Indonesia gradually faces a state of catastrophe. Currently among 726 ethnic languages spoken in Indonesian archipelago, 146 are endangered. Several projects have been initiated for cultural preservation which can prevent the endangered language from being lost. Nevertheless, the available technology that could support communication within indigenous communities, as well as with people outside the community, is still very rare in Indonesia. Speech translation technology is one of the technologies that may help indigenous communities in Indonesia to overcome language barrier and cross cultural gap as well as to face globalization. Our long-term goal is to establish an infrastructure of speech translation system from ethnic languages to English/Indonesian, and this paper presents recent progress of data resources collection and speech recognition system development for four Indonesian major ethnic languages: Javanese, Sundanese, Balinese and Bataks.

Index Terms— Language preservation, Indonesian ethnic languages, speech data collection, speech recognition system.

1. INTRODUCTION

The global, borderless economy and information communication technologies have a great impact in the way of communication. People have to be able to communicate well with others who speak different language. However, on the other hand, intangible cultural expressions, such as oral traditions and literature, are fragile and easily lost. There exist several international projects (i.e., UNESCO's ICT4ID project in 2004-2005) that have been initialized to utilize the use of information and communication technology (ICT) for cultural preservation for preventing them from being lost. Nevertheless, the available technology that could support communication between elders and younger people within indigenous communities, as well as with people outside the community, is still limited. As a result, indigenous communities may still face isolation due to language and cultural barriers.

Indonesia is reported to be one of the most religiously, linguistically, and ethnically diverse regions of the world [1, 2, 3]. It is an archipelago comprising approximately 17500 islands inhabited by hundreds of ethnic groups with more than 241 million people (based on Census 2012). Different ethnic groups speak various different languages. Approximately, there are 300 ethnic groups living in 17,508 islands, that speak 726 native languages [4].

One of the bridges that binds the people together in Indonesia is the usage of *Bahasa Indonesia*, the national language. It is a unity language formed from hundreds of languages spoken in the Indonesian archipelago, which was coined by Indonesian nationalists in 1928 and became a symbol of national identity during the struggle for independence in 1945. Compared to other languages, which have a high density of native speakers, only small proportion of Indonesia's large population speak *Bahasa Indonesia* as a mother tongue while the great majority of people speak it as a second language with varying degrees of proficiency.

Although the phenomena of using the unity language could help the Indonesian people to face the globalization, multilingualism in Indonesia gradually faces a state of catastrophe. Currently among 726 ethnic languages, only thirteen still have a million or more speakers, accounting for 69.91% of the total population, including: Javanese, Sundanese, Malay, Madurese, Minangkabau, Bataks, Bugisnese, Balinese, Acehnese, Sasak, Makasarese, Lampungese, and Rejang [5]. Moreover, of these 13 languages, only 7 languages have presence on the Internet [6]. However, the remaining 713 languages have a total population of only 41.4 million speakers, and the majority of these have very small numbers of speakers [7]. For example, 386 languages are spoken by 5,000 or less; 233 have 1,000 speakers or less; 169 languages have 500 speakers or less; and 52 have 100 or less [8]. These languages are facing various degrees of language

endangerment [9].

The long-term goal is to establish an infrastructure of speech-to-speech translation system from ethnic languages to English/Indonesian. This technology enables communication between the first person who speaks in any language and the other person can understands the meaning of the speech. Therefore, speech translation technology is significant to indigenous communities in Indonesia to overcome language barrier and cross cultural gap as well as to face globalization. This paper presents recent progress of data resources collection and speech recognition system development for four Indonesian major ethnic languages: Javanese, Sundanese, Balinese and Bataks.

In the next section, we briefly describe the overview of standard Indonesian and four major Indonesian ethnic language characteristics. The existing Indonesian resources will be described in Section 3, and the current progress in developing speech corpora for Indonesian ethnic languages will be described in Section 4. Then, Section 5 describes the details of speech recognition system development. Finally, conclusions are drawn in Section 6.

2. INDONESIAN AND ETHNIC LANGUAGES CHARACTERISTICS

2.1. Indonesian National Language

The official/national Indonesian language, so-called Bahasa Indonesia, is a unity language formed from hundreds of languages spoken in the Indonesian archipelago. 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, which was the lingua franca of Southeast Asia[10]. Thus, it is closely related to Malay spoken in Malaysia, Singapore, Brunei, and some other areas. Concerning the number of speakers, today Malay-Indonesian ranks around sixth in size among the world's languages. The only difference is that Indonesia (which was a Dutch colony) adopted the Van Ophuysen orthography in 1901, while Malaysia (which was a British colony) adopted the Wilkinson orthography in 1904. In 1972, the governments of Indonesia and Malaysia agreed to standardize the "improved" spelling, which is now in effect on both sides. Even so, modern Indonesian and modern Malaysian are as different from one another as are Flemish and Dutch.

The standard Indonesian language is continuously being developed and transformed to make it more suitable to the diverse needs of a modernizing society. Many words in the vocabulary reflect the historical influence of various foreign cultures that have passed through the archipelago. It has borrowed heavily from Indian Sanskrit, Chinese, Arabic, Portuguese, Dutch, and English. Although the earliest records in Malay inscriptions are syllable-based written in Arabic script, modern Indonesian is phonetic-based written in Roman script. It use only 26 letters as in the English/Dutch alphabet. The Indonesian phoneme set is defined on the basis of an Indonesian grammar text [11]. A full phoneme set contains a total of 33 phoneme symbols, which consist of 10 vowels (including diphthongs), 22 consonants.

2.2. Ethnic Languages

On the other hand, some of ethnic groups in Indonesia still use their own transcription in daily life. As the four major ethnic groups in Indonesia, Javanese, Sundanese, Balinese and Bataks are counted in that category. Each of these four major Indonesian ethnic languages are further discussed in the following:

1. Javanese
   Javanese language had a long history of its development. Based on the evidence in the form of inscriptions and paleography, the earlier stage of Javanese script was started before the eight century [12]. Javanese transcription is called *Aksara Hanacaraka*. It consists of 20 basic scripts called Carakan, including 20 consonants and 1 vowel. Letter is called *Nglegena* 'naked' because it has not had any *Sandhangan* or clothes that could make them into another vowel sounds. To make Javanese vowels have another sound, it needs an additional tool called *Sandhangan*. Fig. 1(a) shows Javanese script[1]. Currently, Hanacaraka is already included in Unicode (A980-A9DF).

2. Sundanese
   Sundanese has been written in a number of scripts. Pallawa or Pra-Nagari was first used in West Java to write Sanskrit from the fifth to eighth centuries, and from Pallawa was derived Sunda Kuna or Old Sundanese which was used in the Sunda Kingdom from the 14th to 18th centuries [13]. Modern Sunda transcription called *Aksara Sunda*. Aksara means transcription in Indonesia. Similar with Hanacaraka, Aksara Sunda shown in Fig. 1(b) also has basic alphabets, vowels and punctuation to change phoneme and basic punctuation[2]. Basic letters in Aksara Sunda, has also been registered in Unicode (1B80-IBBF).

3. Balinese
   The Balinese script is without doubt derived from Devanagari and Pallava script from India. The shape of the script shows similarities with southern Indian scripts like Tamil. The concept of syllable also found in other

[1] The official site of Aksara Jawa. http://hanacaraka.fateback.com/
[2] http://en.wikipedia.org/wiki/Sundanese_alphabet

South/Southeast Asian scripts, such as the modern Devanagari, Tamil, Thai, Lao, and Khmer scripts. Figure 1(c) shows the Balinese script[3]. The closest sibling is the Javanese script which have rectangular form of font shape compared to round shape of Balinese script [14]. The registered Unicode is 1B00-1B7F.

4. Bataks
   Batak tribe, mainly living in northern region of Sumateran Island (Sumatera Utara) in Indonesia, has been established for around 800-1000 years. Within that long period, Batak people developed several subtribes and clans. The largest one (in population number) is Toba subtribe, followed (in no particular order) by Karo, Simalungun, Pakpak-Dairi, Angkola-Mandailing, and Nias (Niha) people. Batak tribe has its own writing system which existed since 13th century AD. Batak people themselves call their writing system Surat Batak (Surat = letters/writings) [15] shown in Fig. 1(d). Currently, it is already included in Unicode (1BC0-1BFF).

In this preliminary work, to simplify the task we only use the romanized version of those ethnic scripts.

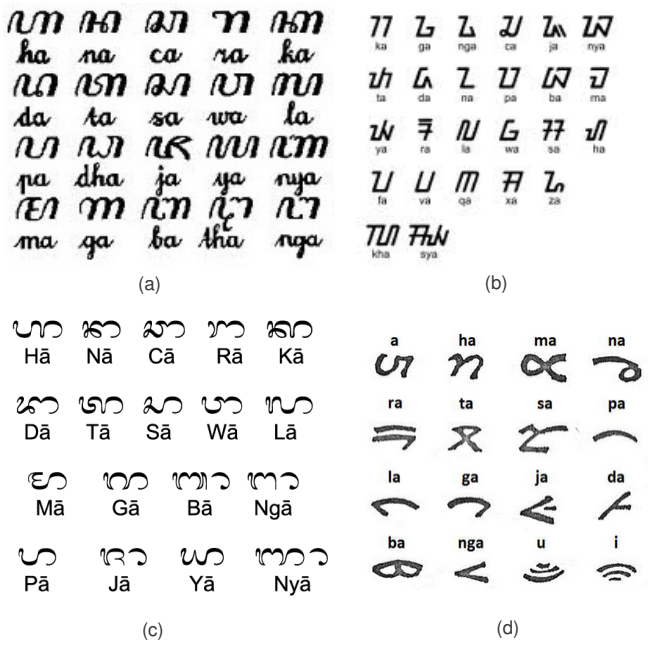

Fig. 1. Script of ethnic languages: (a) aksara Hanacaraka, (b) aksara Sunda, (c) Balinese script, and (d) surat Bataks.

[3]http://en.wikipedia.org/wiki/Balinese_alphabet

3. EXISTING INDONESIAN LARGE VOCABULARY DATA RESOURCES

The Indonesian speech 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 [16, 17]. Two types of Indonesian data resources available in both text and speech forms were used here: daily news task and telephone application task. They are described below.

3.1. Daily News Task

A raw text source for the daily news task has already been generated by an Indonesian student [18]. 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.

From this text data, we then selected phonetically-balanced sentences by using the greedy search algorithm [19]; this produced a total of 3168 sentences. Then, clean and telephone speech utterances 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 original accents were covered: standard Indonesian, Bataks, Javanese, and Sundanese, denoted as "IND", IND-BTK","IND-JAW", and "IND-SND", respectively. However, all speakers were requested to utter 110 sentences without any accents. In total, there were 44,000 speech utterances, which amounted to around 43.35 hours of speech.

3.2. Telephone Application Task

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).

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. DESIGN AND COLLECTION OF INDONESIAN ETHNIC LANGUAGES

4.1. Raw Text Sources Collection

Raw text sources are collected from online newspaper and magazine: Penjabar-Semangat[4] for Javanese, Sunda-News[5] for Sundanese, Bali-Post[6] for Balinese, and Halo-Moantondang[7] for Bataks. Table 4.1 shows the total number of articles and sentences which have successfully been collected.

Table 1. *Raw Text Corpora of Javanese, Sundanese, Balinese and Bataks.*

Languages	# Articles	# Sentences
Javanese	1583	43336
Sundanese	1693	39770
Balinese	3919	20436
Bataks	1096	36204

4.2. Pre-Processing and Validation

The initial forms of these documents contain numbers, punctuation, abbreviations, acronyms, names, and foreign words. We then further processed the raw text sources to generate clean text corpora by:

- converting all upper case letters into lower case
- removing punctuation
- changing numbers into words
- select short sentences (max. 15 words/sentence)

resulting 6616 sentences of Javanese, 5717 sentences of Sundanese, 3249 Sentences of Balinese, and 6870 sentences of Bataks, respectively.

After that, we selected 1000 sentences from cleaned text corpora of each language to be validated by the native speakers. The validation is done in order to correct any errors in the sentences, as well as remove inappropriate sentences, resulting 823 sentences of Javanese, 954 sentences of Sundanese, 956 Sentences of Balinese, and 910 sentences of Bataks, respectively.

4.3. Graphemically-balanced Sentences

Given the validated text corpora, we then selected balanced sentences by using the greedy search algorithm [19]; However, as the phonetically transcription of these ethnic languages are unknown, we selected balanced sentences based on grapheme transcription. This produced a total of 225 sentences for each language as shown in Table 2.

4.4. Parallel Sentences

In addition to graphemically balanced sentences, we also created fifty sentences of Indonesian language based on the ATR basic travel expression corpus (BTEC) which has served as the primary source for developing broad coverage speech translation systems [20]. Those sentences were then translated into Javanese, Sundanese, Balinese, and Bataks languages by native speakers. Table 4.1 shows a translated sentence example of "Apakah anda bersedia untuk makan dengan saya besok malam?" (meaning "Would you like to have dinner with me tomorrow night?").

4.5. Speech Corpora Collection

For speech recording, 40 native speakers were participated. Ten native speakers (5 males and 5 females) of each Javanese, Sundanese, Balinese, Bataks language, which originally came from ethnics of Java, Sunda, Bali, and North Sumatra. Each speaker was asked to utter 325 sentences, including 225 graphemically balanced sentences and 50 parallel sentences (50 Indonesian sentences and 50 ethnic language sentences). All speakers were requested to keep their accent even during uttering Indonesian sentences. The Indonesian utterances with Javanese, Sundanese, Balinese, Bataks accents are denoted as "ACC-JAW", "ACC-SND", "IND-BLI", and "IND-BTK", respectively, while the utterances of Javanese, Sundanese, Balinese, Bataks languages are denoted as "JAW", "SND", "BLI", and "BTK". The speech recording was conducted in a sound proof room in Jakarta, Indonesia. Speech was recorded into WAV file at a 48 kHz sampling rate with 16 bits resolution. The sampling rate was later down-sampled to 16 kHz for further experiments.

5. DEVELOPMENT OF GRAPHEME-BASED SPEECH RECOGNITION SYSTEM

The large vocabulary speech recognition of standard Indonesian language (denoted as "IND") was developed using both daily news and telephone application tasks, while the speech recognition of Indonesian ethnic languages were developed using their specific ethnic language resources, including Javanese, Sundanese, Balinese, Bataks languages (denoted as "JAW", "SND", "BLI", and "BTK"). As the phoneme set of ethnic languages are unknown, all models are develop based on grapheme unit. We also build multilingual acoustic model in which we pooled all data together and build one acoustic model (denoted as "MLT"). The parameter set-up, acoustic modeling, language modeling, pronunciation dictionary and recognition accuracy are described more fully below.

[4] www.penjebarsemangat.co.id
[5] sundanews.com
[6] www.balipost.co.id
[7] halomoantondang.wordpress.com

Table 2. *Number of units and coverage rate of the training data resulting from the greedy search algorithm.*

Graphame	Javanese		Sundanese		Balinese		Bataks	
	# Units	Coverage	# Units	Coverage	# Units	Coverage	# Units	Coverage
Mono-grapheme	27	100%	27	100%	28	100%	23	100%
Left Bi-grapheme	487	86.50%	489	87.79%	441	82.28%	287	90.25%
Right Bi-grapheme	482	86.07%	487	87.59%	438	82.18%	285	90.19%
Tri-grapheme	3269	53.99%	3197	52.56%	2796	53.35%	1767	71.42%

Table 3. *A translated sentence example of "Apakah anda bersedia untuk makan dengan saya besok malam?" (meaning "Would you like to have dinner with me tomorrow night?")*

Languages	Sentences
Indonesian	Apakah anda bersedia untuk makan dengan saya besok malam?
Javanese	Opo kowe gelem mangan bareng aku sesuk bengi?
Sundanese	Dupi anjeun sanggem kanggo tuang sareng abdi enjing wengi?
Balinese	Napikeh mresidayang ragane buin mani peteng ngajeng sareng tiang?
Bataks	Boha molo rap marjobut hita masogot bot ari?

5.1. Front-End Processing

We trained the ASR systems based on weighted finite state transducers (WFSTs) [21] using the Kaldi toolkit [22]. The frontend provides features every 10ms with 25ms width. For each utterances in the speech training data, 13 static mel-frequency cepstral coefficients (MFCC) including zeroth order for each frame are extracted and normalized with cepstrum mean normalization.

To incorporate the temporal structures and dependencies, 9 adjacent (center, 4 left, and 4 right) frames of MFCCs are stacked into one single feature vector leading to 117 dimensional super vectors (9x13 dimensions). It then reduced to an optimum 40 dimensions by applying a linear discriminant analysis (LDA) and maximum likelihood linear transformation (MLLT)[23].

5.2. Model Training

All models are grapheme-based HMM with a standard three-state left-to-right HMM topology without skip states. Each grapheme is classified by its position in word (4 classes: begin, end, internal and singleton). The context-dependent cross-word trigraph HMMs was trained with GMM output probability. This models totally include 15K Gaussians.

The dictionary of standard Indonesian, which is owned R&D TELKOM, was derived from the daily news and telephone application text corpus. It consists of about 40K words in total, including 30K original Indonesian words plus 8K people and place names and 2K foreign words. All pronunciation of these words were manually developed by Indonesian linguists. As for Javanese, Sundanese, Balinese and Bataks ethnic languages, the dictionary was derived from the selected text covering only about 2K words.

Using the SRILM toolkit [24], we built n-gram language models with modified Kneser-Ney smoothing [25] from each of the text corpora (standard Indonesian and ethnic languages). The resulting language model contains 19K trigrams of Indonesian, about 4K trigrams of each ethic language.

5.3. Recognition Accuracy

In this experiments, we investigated the performance of ASR systems on following test sets:

- Standard Indonesian test set ("IND","IND-JAW", "ACC-SND", and "IND-BTK") in which the speakers utter the Indonesian sentences without any accent.
- Accented Indonesian test set ("ACC-JAW", "ACC-SND", "IND-BLI", and "IND-BTK") in which the speakers utter the Indonesian sentences with heavy ethnic accent.
- Ethnic test set ("JAW", "SND", "BLI", and "BTK") in which the speakers utter the sentences of ethnic languages.

First, we investigated the performance of ASR using only standard Indonesian acoustic model on various test sets: standard Indonesian, accented Indonesian and ethnic Indonesian test sets as shown in Fig. 2. The pronunciation dictionary and language model were set-up based on the corresponding test set. The performance of ASR using only Indonesian acoustic model on standard Indonesian test set could achieve less than 1% WER as expected. But then, the performance degrades on speech with heavy accents. Furthermore, it drops significantly up to more than 80% WER on ethnic test sets.

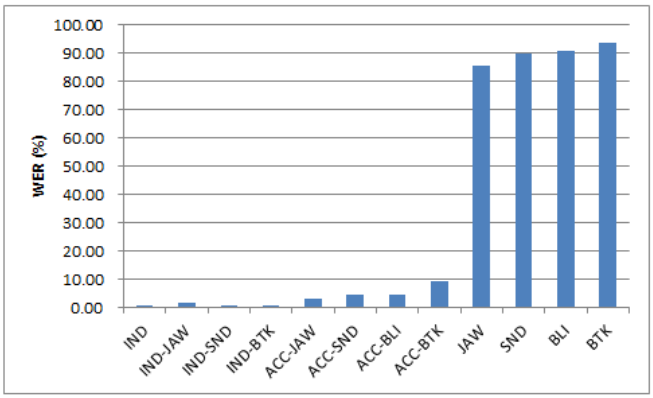

Fig. 2. The performance of standard Indonesian ASR on various test sets: standard Indonesian, accented Indonesian and ethnic Indonesian test sets.

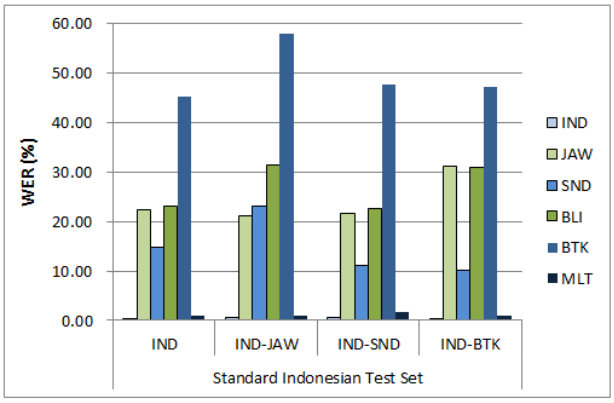

Fig. 3. The performance of various ASR (IND, JAW, SND, BTK, BLI, MLT) on standard Indonesian test set.

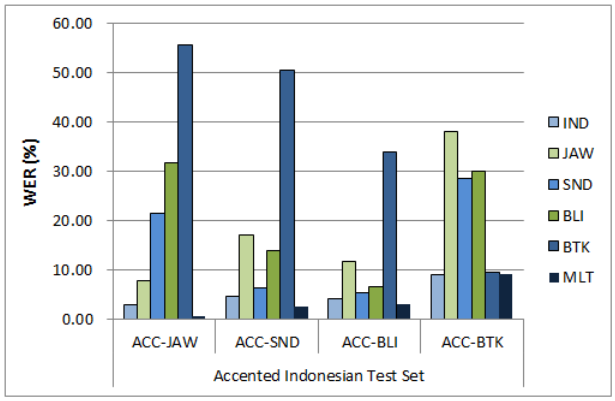

Fig. 4. The performance of various ASR (IND, JAW, SND, BTK, BLI, MLT) on accented Indonesian test set.

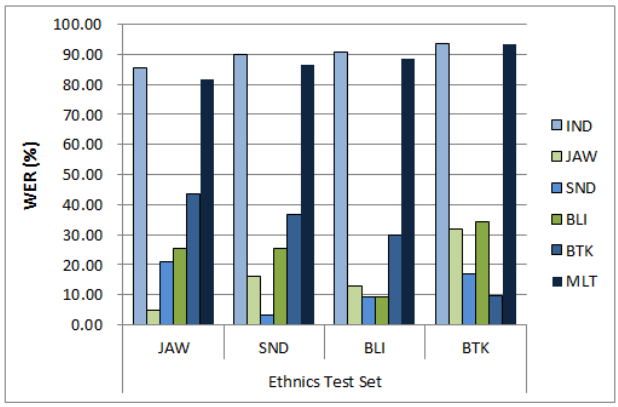

Fig. 5. The performance of various ASR (IND, JAW, SND, BTK, BLI, MLT) on ethnic test set.

Next, Fig. 3, 4, and 5 show the comparison performance between ASR systems that use standard Indonesian acoustic model ("IND"), ethnic-language-specific acoustic model ("JAW", "SND", "BTK", "BLI") and multilingual acoustic model ("MLT"), which were tested on standard Indonesian, accented Indonesian and ethnic Indonesian test sets, respectively. On standard Indonesian test set, the ASR with Indonesian acoustic model perform the best, while on accented Indonesian test set, the ASR with multilingual acoustic model perform the best. These results reveal that combining ethnic speech data could improve the performance on heavily accented of Indonesian speech. However, the performance of on ethnic test set is still very low, which is beyond 80% of WER. In this case, the ASR systems with ethnic-language-specific acoustic model still provide a better performance. The best performances on ethnic test set were still achieved by ethnic-language-matched acoustic model. This may indicate that the acoustic characteristics of Indonesian are quite different from the acoustic characteristic of those ethnic languages. Therefore, finding optimum way to combine various different acoustic speech may be necessary.

6. CONCLUSION

We have presented the collection of Indonesian ethnic speech corpus, which includes Javanese, Sundanese, Balinese, and Bataks spoken languages. This includes 225 graphemically balanced sentences and 50 parallel sentences. We have also presented the current state of ASR system development for Indonesian ethnic languages. The results reveal that the ASR system with multilingual acoustic model could improve the performance on heavily accented of Indonesian speech, but not on ethnic language speech. This may indicate that the acoustic characteristics of Indonesian are quite different from the acoustic characteristic of those ethnic languages. The best performances on ethnic test set were still achieved by ethnic-language-matched acoustic model. In the future, we

will investigate other techniques to combine various different speech acoustic of those ethnic languages, such the use of multilingual multilayer perceptron.

Acknowledgment

Part of this work was supported by Grant-in-Aid for Young Scientists (KAKENHI Wakate-B Grant Number 24700172).

7. REFERENCES

[1] H. Abas, *Indonesian as a unifying language of wider communication: A historical and sociolinguistic perspective*, Pacific Linguistics, Canberra, Australia, 1987.

[2] J. Bertrand, *Language policy in Indonesia: The promotion of a national language amidst ethnic diversity. In Fighting words: Language policy and ethnic relations in Asia*, The MIT Press, Cambridge, MA, USA, 2003.

[3] C.-Y. Hoon, "Assimilation, multiculturalism, hybridity: The dilemas of the ethnic Chinese in post-Suharto Indonesia," *Asian Studies Review*, vol. 7, no. 2, pp. 149–166, 2006.

[4] J. Tan, "Bahasa Indonesia: Between faqs and facts," http://www.indotransnet.com/article1.html.

[5] M. Lauder, *Language Treasures in Indonesia. In Words and Worlds : World Languages Review*, Prentice Hall, Clevedon, England, 2005.

[6] Y. MIkami H. Riza, Moedjiono, "Indonesian languages diversity on the internet," in *Internet Governance Forum (IGF)*, Athens, Greece, 2006.

[7] H. Riza, "Indigenous languages of Indonesia: Creating language resources for language preservation," in *Proc. of IJCNLP Workshop on NLP for Less Privileged Languages*, Hyderabad, India, 2008.

[8] G.R. Gordon, *Ethnologue: Languages of the World*, SIL International, Dallas, Texas, USA, 2005.

[9] D. Crystal, *Language Death*, Cambridge University Press, Cambridge, UK, 2000.

[10] G. Quinn, "The Indonesian language," https://www.google.com/#q=Indonesian+language+quinn.

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

[12] H. Kahler and J.G. de Casparis, *Indonesian Paleography: A History of Writing in Indonesia from the Beginning to AD 1500*, E. J. Brill, Leiden/Koln, 1975.

[13] M. Everson, "Preliminary proposal for encoding additional Sundanese characters for old sundanese in the UCS," http://www.unicode.org/L2/L2009/09190-n3648-sundanese.pdf, 2009.

[14] I.B.A. Sudewa, "Contemporary use of the Balinese script," http://www.unicode.org/L2/L2003/03118-balinese.pdf, 2003.

[15] A. Samosir, "Surat Batak," http://www.ancientscripts.com.

[16] S. Sakti, E. Kelana, H. Riza, S. Sakai, K. Markov, and S. Nakamura, "Recent progress in developing Indonesian large-vocabulary corpora and LVCSR system," in *Proc. MALINDO*, Cyberjaya-Selangor, Malaysia, 2008, pp. 40–45.

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

[18] 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.

[19] 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.

[20] 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.

[21] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech and Language*, vol. 20, no. 1, pp. 69–88, 2002.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Moticek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, Hawaii, USA, 2011.

[23] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[24] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. of ICSLP*, Denver, USA, 2002, pp. 901–904.

[25] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Proc. ICASSP*, 1995, pp. 181–184.