Towards Language Preservation: Design and Collection of Graphemically Balanced and Parallel Speech Corpora of Indonesian Ethnic Languages

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*—Various intangible cultural expressions in Indonesia such as oral traditions and literature are fragile and easily lost. Currently among 726 languages, 146 are endangered. Although several projects have been initiated for cultural preservation, 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-to-speech translation is a technology that enables communication among people speaking in different languages, and therefore it is significant for indigenous communities to preserve their cultural language and overcome language barriers. This paper presents the earlier step of long-term development of speech-to-speech translation system from Indonesian ethnic languages to other languages (i.e., English/Indonesian), which is a design and collection of graphemically balanced and parallel speech corpora of four Indonesian major ethnic languages: Javanese, Sundanese, Balinese and Bataks.**

I. Introduction

Cultural diversity helps make our world rich and vital. The world's population of indigenous people now numbers some 350 million individuals representing nearly 6000 languages and cultures. However, with the advent of globalization, intangible cultural expressions, such as oral traditions and literature, are fragile and easily lost. Indonesia is reported to be one of the most religiously, linguistically, and ethnically diverse regions of the world [1]–[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. Table I show thirteen of the indigenous ethnic languages that 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]. 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].

TABLE I: *Thirteen of the indigenous ethnic languages in Indonesia which still have a million or more speakers [5].*

Languages	# Speakers
Javanese	75,200,000
Sundanese	27,000,000
Malay	20,000,000
Madurese	13,694,000
Minangkabau	6,500,000
Batak	5,150,000
Bugisnese	4,000,000
Balinese	3,800,000
Acehnese	3,000,000
Sasak	2,100,000
Makasarese	1,600,000
Lampungese	1,500,000
Rejang	1,000,000

There exists 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.

Speech-to-Speech (S2S) translation is a technology that translates spoken language into speech in another language. It enables communication with people speaking in different languages, and therefore speech-translation technology is significant for indigenous communities to overcome language barriers and cross-cultural gap. This paper presents the earlier step of long-term development of speech-to-speech translation system from Indonesian ethnic languages to other languages (i.e., English/Indonesian), which is a design and collection of graphemically balanced and parallel speech corpora of four major Indonesian 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 design of graphemically balanced text database will be described in Section III, and the development of speech corpora will be described in Section IV. Then, Section V describes the analysis of pitch range difference in Indonesian and native ethnic languages. Finally, we draw our summaries in Section VI.

II. Indonesian Ethnic Languages Characteristics

The official Indonesian language, so-called Bahasa Indonesia, is a unity language formed from hundreds of languages spoken in the Indonesian archipelago. It is basically derived from the literary of the Malay dialect, which was the lingua franca of Southeast Asia [10]. In earliest records, Malay inscriptions are syllable-based written in Arabic script, however modern Indonesian is currently phonetic-based written in Roman script. It uses only 26 letters as with the case of the English/Dutch alphabet.

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 script 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 [11]. Javanese transcription is called *Aksara Hanacaraka*. It consists of 20 basic scripts called Carakan, including 20 consonants and 1 vowels. The 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 shows Javanese script *Hanacaraka*[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

Fig. 1: Javanese script.

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 [12]. Modern Sunda transcription called *Aksara Sunda*. Aksara means transcription in Indonesia. Similar with Hanacaraka, Aksara Sunda shown in Fig. 2 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).

Fig. 2: Sundanese script.

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 South/Southeast Asian scripts, such as the

[1]The official site of Aksara Jawa. http://hanacaraka.fateback.com/

[2]http://en.wikipedia.org/wiki/Sundanese_alphabet

modern Devanagari, Tamil, Thai, Lao, and Khmer scripts. Figure 3 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 [13]. The registered Unicode is 1B00-1B7F.

Fig. 3: Balinese script.

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) [14] shown in Fig. 4. Currently, it is already included in Unicode (1BC0-1BFF).

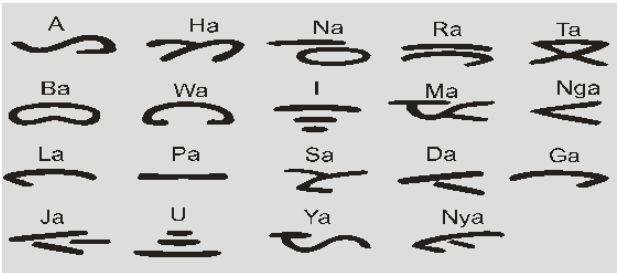

Fig. 4: Bataks script.

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

III. Database Design

A. Pre-Processing and Validation

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 II shows the total number of articles and sentences which have successfully been collected.

TABLE II: *Raw Text Corpora of Javanese, Sundanese, Balinese and Bataks.*

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

2) Text Preprocessing

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 for each sentence)

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

3) Text Validation by Native Speakers

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 grammatical errors in the sentences, as well as remove inappropriate or impolite sentences, resulting 823 sentences of Javanese, 954 sentences of Sundanese, 956 Sentences of Balinese, and 910 sentences of Bataks, respectively.

B. Graphemically-balanced Sentences

Given the validated text corpora, we then selected balanced sentences by using the greedy search algorithm [15]; However, the challenge occurs since the phonetically transcription of these ethnic languages are unknown. In Indonesian language, all letters are pronounced much more consistently, and no letters are muted. Although some exceptions exist, there is

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

TABLE III: *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 IV: *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?

TABLE V: *Pitch range in Hz of four speakers in Indonesian (L2) and their native ethnic languages (L1).*

Gender	Speaker	F0 Max		F0 Min		F0 Range	
		L1	L2	L1	L2	L1	L2
Male	Javanese	242.26	230.73	73.57	76.05	168.69	154.68
	Sundanese	247.67	238.75	91.37	96.31	156.30	142.44
	Balinese	315.16	305.91	82.84	84.61	232.32	221.30
	Bataks	265.24	246.03	47.66	52.05	217.58	193.98
Female	Javanese	367.33	360.60	78.36	79.96	288.97	280.64
	Sundanese	438.24	437.66	116.97	123.39	321.27	314.27
	Balinese	352.03	322.81	88.47	91.46	263.56	231.35
	Bataks	429.66	407.90	75.75	74.02	353.91	333.88

fairly good match between spelling and pronunciation. Based on the assumption that Javanese, Sundanese, Balinese and Bataks languages may have similar condition, we proposed to select balanced sentences based on grapheme transcription. This produced a total of 225 sentences for each language as shown in Table III.

C. 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 [16]. Those sentences were then translated into Javanese, Sundanese, Balinese, and Bataks languages by native speakers. Table IV 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?").

IV. 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). 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.

Because the texts were taken from four different ethnic languages, file names have to be discriminated, as follows $EEEXXX\_F/M\_L\_C\_news\_YYYY.wav$ where:

- EEE is the code for ethnic languages, "Jaw" for Javanese, "Snd" for Sundanese, "Bli" for Balinese, and "Btk" for Bataks,
- XXX is the order of speaker (in this matter 001-010),
- F is female speaker and M for male speaker,
- L is another code for ethnic languages, J for Java and S for Sunda,
- C_news means this speech built by reading news text clearly, and
- YYYY is the order of speech.

V. Pitch Range in L1 (Ethnic) and L2 (Indonesian)

Research indicates that languages may differ in how pitch range is manifested [17]. Here, we investigate whether the

pitch range is different when a person speaks in Indonesian and their ethnic languages. This analysis process is based on fundamental frequency extraction using TEMPO (Time-Domain Excitation extractor using Minimum Perturbation Operator) [18]. The pitch range of four speakers (one for each ethnic language) varied according to the language they speak is shown in Table V. All speakers had a wider pitch range in their native ethnic languages than in Indonesian.

VI. Summary

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. In future direction, we will utilize these ethnic language corpora and study how to build a speech-to-speech translation System for Indonesian ethnic languages in rapid way by the use of existing Indonesian speech recognition.

Acknowledgment

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

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. 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.

[12] 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.

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

[14] A. Samosir, "Surat batak," http://www.ancientscripts.com.

[15] 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.

[16] 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.

[17] R. Van Bezooijen, "Sociocultural aspects of pitch differences between japanese and dutch women," *Language and Speech*, vol. 38, no. 3, pp. 253–265, 1995.

[18] H. Kawahara, H. Katayose, A. de Cheveigne, and R. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and perodicity," in *Proc. EUROSPEECH*, Budapest, Hungary, 1999, pp. 2781–2784.