

TOWARDS LANGUAGE PRESERVATION: PRELIMINARY COLLECTION AND VOWEL ANALYSIS OF INDONESIAN ETHNIC SPEECH DATA

Auliya Sani^{1,2}, Sakriani Sakti¹, Graham Neubig¹, Tomoki Toda¹, Adi Mulyanto², Satoshi Nakamura¹

¹Nara Institute of Science and Technology, Japan

²Bandung Institute of Technology, Indonesia

{aulya-f, ssakti, neubig, tomoki, s-nakamura}@is.naist.jp

{13509067@std, adi@}.stei.itb.ac.id

ABSTRACT

Multilingualism in Indonesia gradually faces a state of catastrophe. Although several projects have been initiated for cultural preservation, 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 very rare in Indonesia. This paper presents the first step of long-term development of speech-to-speech translation system from ethnic languages to English/Indonesian, which is collection and analysis of Indonesian ethnic speech corpora. Here, we will first focus on two largest ethnic groups in Indonesia which are Javanese and Sundanese.

Index Terms— Language preservation, Indonesian ethnic languages, speech data collection, vowel analysis

1. INTRODUCTION

Indonesia is an archipelago comprising approximately 17500 islands inhabited by hundreds of ethnic groups with more than 241 million people (based on Census 2012). Two largest ethnic groups are Javanese and Sundanese living in Java Island. Different ethnic groups speak various different languages. One of the bridges that binds the people together 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. To promote the usage of Indonesian language, the government makes a strong campaign to use Indonesian in daily life.

On the other hand, 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. As an international language, English became the most spoken language in the world with more than 1800 million speakers. Thus, in modern Indonesia, along with the campaign of *Bahasa Indonesia*, English is also promoted starting at primary education.

Although the phenomena of using the unity language, such as Indonesian as official Indonesia language or English as a world language could help the Indonesian people to face the globalization, multilingualism in Indonesia gradually faces a state of catastrophe. Currently from 726 languages, 146 are endangered. An endangered language is a language that is at risk of falling out of use, generally because it has few surviving speakers. If it loses all of its native speakers, it becomes an extinct language. In near future, more and more languages will be endangered.

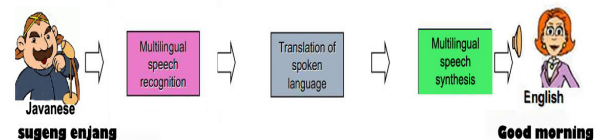


Fig. 1. Speech translation system mechanism between ethnic languages to English/Indonesian.

Several projects have been initiated for cultural preservation which can prevent the endangered language from being lost such as holding language congress, documenting the words, making rules for public servant to speak in ethnic language in a given day, etc. 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 very rare in Indonesia. As a result, indigenous communities may still face isolation due to language and cultural bar-

rier. The long-term goal is to establish an infrastructure of speech-to-speech translation system from ethnic languages to English/Indonesian (See Fig. 1). 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 indigenious communities in Indonesia to overcome language barrier and cross cultural gap as well as to face globalization.

This paper presents the first step of developing speech technology, which is collection and analysis of Indonesian ethnic speech corpora. Here, we will first focus on two largest ethnic groups in Indonesia which are Javanese and Sundanese. In the next section, we briefly describe the overview of Javanese and Sundanese languages. The existing Indonesian data will be described in Section 3, and the current development of Javanese and Sundanese speech corpus will be described in Section 4. In Section 5, the analysis of vowels in standard Indonesian, Javanese, and Sundanese is presented. Finally, we draw our conclusions in Section 6.

2. JAVANESE AND SUNDANESE LANGUAGES CHARACTERISTICS

2.1. Script of Languages

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. 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 two largest ethnic groups in Indonesia, Javanese and Sundanese are counted in that category. Even in elementary school education, the subject of learning this language is still given. Javanese transcription is called *Aksara Hanacaraka* and Sunda transcription called *Aksara Sunda*. *Aksara* means transcription in Indonesia. These transcriptions derive from Pallawa transcription from South India. Figure 2 shows Javanese script *Hanacaraka*. It consists of 20 basic alphabets called *Carakan*, 20 consonants, 5 vowels, transcription for numbers and for foreign words, the honorifics and punctuation called *Sandhangan*. To make a different syllable, we have to add *Sandhangan* to change the phoneme. Similar with *Hanacaraka*, *Aksara Sunda* shown in Fig. 3 also has basic alphabets, vowels and punctuation to change phoneme and basic punctuation.



Fig. 2. Javanese script.



Fig. 3. Sundanese script.

2.2. Phoneme Set

Indonesian phoneme set consist of 10 vowels (including diphthongs) and 22 consonants [1]. It consist of vowel, i.e., /a/ (like "a" in "father"), /i/ (like "ee" in "knee"), /u/ (like "oo" in "moon"), /e/ (like "e" in "bed"), /ə/ (a schwa sound, like "e" in "lantern"), /o/ (like "o" in boss), and four diphthongs, /ay/, /aw/, /oy/, and /ey/.

Javanese and Sundanese phoneme sets are similar with those of Indonesian. Sundanese phoneme set contains 7 vowels and 21 consonants. The difference is that Sundanese has no diphthong in vowel, but instead it has another vowel called /eu/. Basically, Sundanese does not also have /f/ and /z/ in consonants. However, as influence by the use of Indonesian language, nowadays Sundanese covers these consonants as well. But still, Sundanese does not have /kh/ and /sy/ as in Indonesian. Similar with Sundanese, Javanese phoneme set has no diphthongs and did not have /f/ and /z/ previously. In addition, Javanese has /ɛ/ and /ɔ/ (like "a" in saw). In total, Javanese phoneme set contains 8 vowels and 23 consonants.

Unlike Indonesian and Sundanese, Javanese has many rules in reading especially for vowel. For example, when alphabet 'a' lies in the end of word, it is called open syllable and sometimes spoken as /ɔ/. If 'a' meets consonant (closed syllable), it is spoken as /a/. The vowel articulation pattern indicates the first two resonances of vocal tract, F1 for height and F2 for backness. The comparison of Indonesian, Javanese, and Sundanese vowel articulation pattern is shown

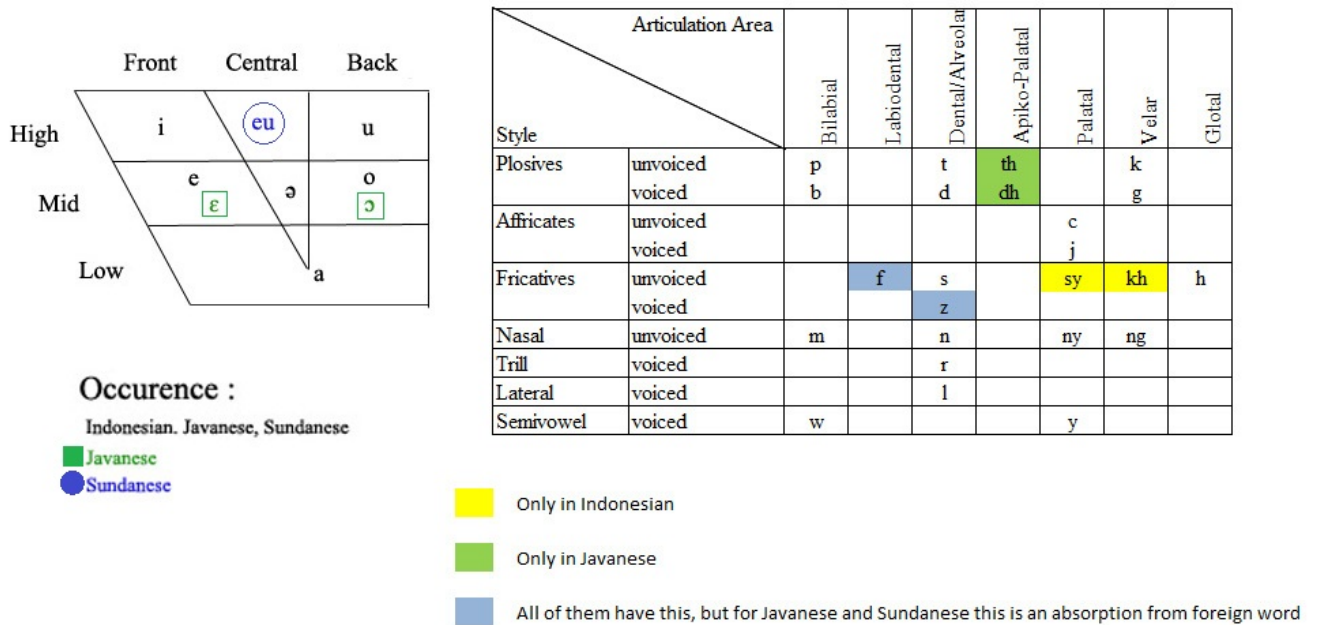


Fig. 4. Articulatory pattern for vowels and consonants of Javanese, Sundanese, and Standard Indonesian.

in Fig. 4 (left side). Consonants can be made by changing articulation area, style, and also vocal cords condition. The comparison of Indonesian, Javanese, and Sundanese's consonant articulatory pattern can be seen in Fig. 4 (right side).

3. EXISTING INDONESIAN 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 [2, 3].

A raw text source for the daily news task has already been generated by an Indonesian student [4]. 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 [5]; 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 main accents were covered: Bataks, Javanese, Sundanese, and standard Indonesian (without accent). Each speaker uttered 110 sentences, resulting in a total of 44,000 speech utterances, which

amounted to around 43.35 hours of speech.

4. COLLECTION OF JAVANESE AND SUNDANESE SPEECH DATA

4.1. Text Corpus

Two documents are collected from newspaper, Mangle¹ online collection for Sundanese and Djaka Lodang magazine for Javanese. The initial forms of these documents contain numbers, punctuation, abbreviations, acronyms, names, and foreign words. These documents were then converted to another documents by:

- converting all upper case letters into lower case
- removing punctuation
- changing numbers into words
- select short sentences (max. 20 words for each sentence)

From this text data, we then selected phonetically-balanced sentences by using the greedy search algorithm[5]; this produced a total of 230 sentences as shown in Table 1.

4.2. Speech Recording

For recording, 20 native speakers were selected, 10 native speakers (5 males and 5 females) from Java ethnic and the other 10 native speakers (5 males and 5 females) from Sunda

¹majalah-mangle.com

Table 1. Javanese and Sundanese Text Corpora

Attributes	Javanese	Sundanese
Number of Sentence	230	230
Number of Words	2999	4262
Vocabulary Size	1529	1728
Number of Name Words	24	19
Number of Foreign Words	2	3

ethnic. Each speaker was asked to read prepared text of the 230 sentences. Speech was recorded in two different places, in Indonesia and Japan. Speech was recorded in a quiet room. Recording materials were a Sony ECM-674 and a Sennheiser HMD 280 Pro. Speech was recorded into WAV file at 44.1 kHz sampling frequency 16 bit (sample size). Because the texts were taken from two different ethnic languages, file names have to be discriminated. For Sundanese speech file, the label is *EEEXXX_F/M_L_C_news.YYYY.wav* where:

- EEE is the code for ethnic languages, Jaw for Java and Snd fo Sunda,
- 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.

5. VOWEL ANALYSIS OF JAVANESE AND SUNDANESE

As described in Section 2.2, there is a difference between vowel articulatory pattern in Indonesian, Javanese, and Sundanese. This section describes the experiment in comparing Indonesian, Javanese, and Sundanese vowel based on the data that has been collected. Here, Praat² tools are used to create formant charts. Two speakers (female and male) with heavily accents were chosen from each language. Then, the segmented phoneme data from those female speaker and male speaker was processed separately. Here, we selected only the phoneme with previous nasal context which is commonly used in these languages. After normalize the F1 and F2, the formant charts were obtained (similar with experiments described in [6]). In all charts, Indonesian vowels from experiment are marked by black font while Javanese vowel are marked by blue font. In the first two charts, Sundanese vowels are marked in purple.

²praat.en.softonic.com/

First, we examine /e/, /ə/, /ɛ/ of Javanese vowels and /e/, /ə/, /eu/ in Sundanese vowels. In comparison, we include also vowel /e/ and /ə/ from standard Indonesian. The first chart (shown in Fig. 5) was obtained from female speaker data and the second chart (shown in Fig. 6) from male. These charts show vowels' location in F1 and F2 scale in Hertz. From both charts we can see the difference in vowels' location from each language. The lower the F1 value, the higher (closer) the vowel. For both /e/ and /ə/, Sundanese vowels lie on the top of other languages, followed by Javanese vowel and Indonesian vowel at the bottom. Even for the same vowels, the formant values of each language are located in slightly different area. The reasonable explanation of this phenomenon is the style of speaking from each ethnic group itself. They use different dialect in speaking (further details can also be found in [7]). Furthermore, the position of Javanese /ɛ/ lies in middle position closed to Indonesian /e/, while the position of Sundanese /eu/ lies in central-high position. But if we look entirely, the pattern from vowels is similar to articulatory pattern for vowels in Fig. 4.

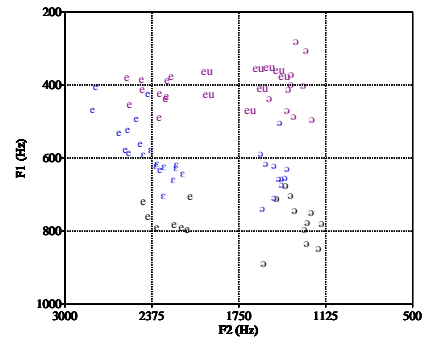


Fig. 5. Vowels distribution of /e/, /ə/, /ɛ/, and /eu/ depending on F1 and F2 for Javanese, Sundanese, and Indonesian female speakers.

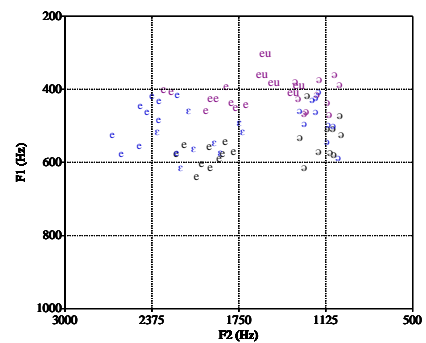


Fig. 6. Vowels distribution of /e/, /ə/, /ɛ/, and /eu/ depending on F1 and F2 for Javanese, Sundanese, and Indonesian male speakers.

Next, we examine /a/, /o/, /ɔ/ of Javanese vowels, in comparison with vowel /a/ and /o/ from standard Indonesian. The

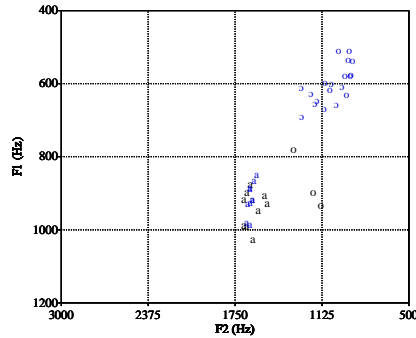


Fig. 7. Vowels distribution of /a/, /o/, and /ɔ/ depending on F1 and F2 for Javanese and Indonesian female speakers.

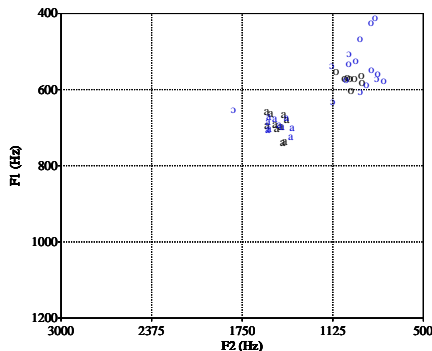


Fig. 8. Vowels distribution of /a/, /o/, and /ɔ/ depending on F1 and F2 for Javanese and Indonesian male speakers.

first chart (shown in Fig. 7) was obtained from female speaker data and the second chart (shown in Fig. 8) from male. Similar tendencies as before, both charts show that the formant values of /o/ from Javanese are generally lower, meaning that the vowels lie higher than those of Indonesian speech. On the other hand, the position of Javanese /ɔ/ lies in middle position closed to Indonesian /o/. However, the formant values of /a/ from Javanese and standard Indonesian lie almost in the same region.

Comparing the charts obtained from female and male speakers, the positions of vowels are different. However, anatomically-related acoustic differences in the speech of males and females are well-established in the literature [8, 9]. The vocal tract of adult females is shorter than that of adult males, causing higher F0 and higher formant frequencies.

6. CONCLUSION

We have presented the development of Indonesian ethnic speech corpus, which includes Javanese and Sundanese languages. An experiment to analyze the vowels between Javanese, Sundanese and standard Indonesian has also been done. The results reveal that even for the same vowels,

the formant values of each language are located in slightly different area. In other words, the vowels in Javanese and Sundanese are acoustically higher than standard Indonesia. 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.

7. REFERENCES

- [1] H. Alwi, S. Dardjowidjojo, H. Lapoliwa, and A.M. Moeliono, *Tata Bahasa Baku Bahasa Indonesia (Indonesian Grammar)*, Balai Pustaka, Jakarta, Indonesia, 2003.
- [2] 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.
- [3] 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.
- [4] 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.
- [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] H. Wang and V.J. Van Heuven, "Acoustical analysis of english vowels produced by chinese, dutch, and american speakers," 2006, pp. 237–248.
- [7] E.V. Zanten and V.J.V. Heuven, "The indonesian vowels as pronounced and perceived by toba batak, sundanese and javanese speaker," *Bijdragen tot de Taal-, Land- en Volkenkunde*, vol. 140, no. 4, pp. 497–521, 1984.
- [8] G. Fant, "A note on vocal tract size factors and non-uniform f-pattern scalings," *STL-QPSR*, vol. 7, no. 4, pp. 22–30, 1966.
- [9] R.L. Diehl, B. Lindblom, K.A. Hoemeke, and R.P. Fahy, "On explaining certain male-female differences in the phonetic realization of vowel categories," *Journal of Phonetics*, vol. 24, pp. 187–208, 1996.