

# Indonesian Phoneme Set, Vocabulary, and Pronunciation for Automatic Speech Recognition and Speech Synthesizer

Dessi Puji Lestari<sup>1,2</sup>, Roland Hartanto<sup>1</sup>, Devin Hoesen<sup>2</sup>,  
Guntario Sukma Cahyani<sup>2</sup>, Sakriani Sakti<sup>3§</sup>

<sup>1</sup>Institut Teknologi Bandung, Jl. Ganesha 10 Bandung, Indonesia

<sup>2</sup>Prosa.ai, Jl. dr. Otten 10 Bandung, Indonesia

<sup>3</sup>Nara Institute of Science and Technology, 8916-5 Takayama, Nara 630-0192, Japan

dessipuji@stei.itb.ac.id, 13515107@std.stei.itb.ac.id, {devin.hoesen, guntario.cahyani}@prosa.ai, ssakti@is.naist.jp

## Abstract

This paper describes the design of the Indonesian phoneme set, vocabulary and pronunciation applied for two main speech technology applications, automatic speech recognition and automatic speech synthesizer research for Bahasa Indonesia. There are 32 standard Indonesian phonemes, but they do not include variations in sound pronunciation or allophones caused by dialect and foreign language influences. Some studies of the Indonesian speech recognition systems even try to reduce the standard phoneme class, especially sounds that caused by the influence of foreign languages that are difficult to pronounce by Indonesians and diphthong sounds to improve the accuracy of speech recognition systems. On the other hand, from the results of the speech synthesizing system it was found that 32 standard Indonesian phonemes without incorporating allophones were insufficient to represent the pronunciation of Indonesian, thus the synthesized speech was not as natural as the original speaker.

**Keywords:** Bahasa Indonesia, phoneme, vocabulary, pronunciation, speech recognition, speech synthesizer

## Abstrak

Pada makalah ini dipaparkan rancangan set bunyi atau fonem, kosakata, dan pelafalan kata yang digunakan pada penelitian-penelitian sistem pengenalan ucapan dan sistem pensintesis ucapan untuk Bahasa Indonesia. Terdapat 32 bunyi baku bahasa Indonesia, namun bunyi-bunyi tersebut belum mencakup variasi-variasi pelafalan bunyi atau alofon yang disebabkan oleh pengaruh dialek dan pengaruh bahasa asing. Beberapa penelitian sistem pengenalan ucapan bahasa Indonesia bahkan berusaha mengurangi kelas bunyi standard tersebut terutama bunyi-bunyi yang disebabkan oleh pengaruh bahasa asing yang sulit diucapkan oleh orang Indonesia dan bunyi diftong untuk meningkatkan akurasi sistem pengenalan ucapan. Sebaliknya, dari hasil penelitian sistem penyintesis ucapan didapat bahwa 32 bunyi baku bahasa Indonesia tanpa memasukkan alofon dinilai masih kurang mewakili pelafalan bahasa Indonesia, sehingga suara hasil sintesis terdengar tidak sealami pembicara aslinya.

## 1. Introduction

Malay language is the root of Indonesian language called Bahasa Indonesia. For a long time, Malay language had indeed been used as an intermediary language or social language in Indonesia archipelago, Brunei, and Malaysia. This language spread very quickly in almost all regions in Indonesia since the 7th century and even formed a separate language variant that differed from its root, called "van Ophuijsen Malay" (Hasan, 1999). This language became widely known among indigenous people and was to be the national identity of Indonesia. It was then formalized as the language of unity in 1928 Youth Congress.

Aside from using Bahasa Indonesia, most Indonesian people also speak their respective regional languages. Indonesia has the most variety of languages in the world after Papua New Guinea. There are 707 languages that are used as a first language by the Indonesian population (Lewis et al., 2013). According to the Language Development Agency, Ministry of Education and Culture, there were 668 regional languages in Indonesia in 2018.

Since the regional languages in the eastern region has not yet been fully identified, this number is expected to continue to increase. The use of regional languages as a first language for the majority of Indonesian population then affects the pronunciation in various regions in Indonesia and affects vocabulary of Bahasa Indonesia.

In addition to being influenced by regional languages, Bahasa Indonesia is also being influenced by foreign languages through trade and religious missions since before the 4th century. These languages include Sanskrit, Tamil and Hindi from India, Chinese, Arabic, Portuguese, Dutch, and English. Their influence can be seen from the large number of Indonesian words originating from foreign languages (Pastika, 2012).

In developing speech technology for Bahasa Indonesia, especially an automatic speech recognition (ASR) and a speech synthesizer, the influences of foreign languages and local languages on the Indonesian language must be considered especially when designing the phoneme set and the vocabulary list and its pronunciation (lexicon). The next section of this paper will explain Indonesian

<sup>§</sup>This work was done while the author was a member of ATR/NICT Spoken Language Communication Research Labs, Japan

phonemes and vocabulary from the linguistic perspective, and how they are used in some research on automatic speech recognition (ASR) systems, and speech synthesizer or text-to-speech (TTS) systems.

## 2. Bahasa Indonesia

### 2.1 Grapheme Set

The alphabet used in Indonesian spelling consists of 26 letters; ⟨a⟩ until ⟨z⟩ (PUEBI, 2016). It has a highly phonemic orthography, i.e. almost all graphemes represent one phoneme sound, except for a few sounds represented by digraphs and vice versa, almost all phonemes are represented by either one or two graphemes. Exceptions to this rule are found in three phonemes, which are phonemes /e/ and /ə/ both of which are written with the letter ⟨e⟩ or ⟨E⟩, and phonemes /ʔ/ (glottal stop) which are sometimes written with the letter < k > or not written at all. Other exceptions are in absorption words from foreign languages (Yap, 2010).

### 2.2 Phoneme Set

According to Soderberg (2008), Bahasa Indonesia has 32 phonemes: vowels (/a/, /e/, /ə/, /i/, /o/, /u/); diphthongs (/ai/, /au/, /oi/); plosives (/b/, /d/, /g/, /k/, /ʔ/, /p/, /t/); affricates (/tʃ/, /dʒ/); nasals (/m/, /n/, /ŋ/); trill (/r/); fricatives (/f/, /h/, /x/, /s/, /ʃ/, /z/); approximants (/w/, /j/); and lateral approximant (/l/). These phonemes are the standard phonemes used by Indonesians when uttering Indonesian words without considering their allophone.

Sometimes foreign words are also found in daily conversation, especially English words. Because of the foreign language influence, code-switching phenomenon is a common thing found in Bahasa Indonesia. However, most Indonesians pronounce English words using the Indonesian accent, although it varies depending on their English fluency. To cope with the code-switching phenomena, the phoneme set designed for speech technology for Bahasa Indonesia must include the foreign language phoneme analysis. In the CMU (Carnegie Mellon University) pronunciation dictionary there are 39 English phonemes, and some of the phonemes are overlapped with Bahasa Indonesia's phonemes. To determine the phoneme set of English spoken by Indonesian, one of the approaches is by mapping the English phonemes to the most similar phonemes in Bahasa Indonesia as conducted in some speech recognition research for Bahasa Indonesia (Lestari, 2010; Hartanto, 2019). From Lestari (2010), the mappings are as follows (English → Indonesian): /a/→/a/; /æ/→/e/; /ɛ/→/e/; /ʌ/→/ə/; /ɔ/→/o/; /ɑʊ/→/au/; /aɪ/→/ai/; /b/→/b/; /tʃ/→/tʃ/; /d/→/d/; /ð/→/d/; /ɜ/→/e/ /ɪ/; /eɪ/→/e/ /j/; /f/→/f/; /g/→/g/; /h/→/h/; /I/→/i/; /i/→/i/; /dʒ/→/dʒ/; /k/→/k/; /l/→/l/; /m/→/m/; /n/→/n/; /ŋ/→/ŋ/; /oʊ/→/o/; /ɔɪ/→/o/; /p/→/p/; /ɪ/→/ɪ/; /s/→/s/; /ʃ/→/ʃ/; /t/→/t/; /θ/→/t/; /ʊ/→/u/; /u/→/u/; /v/→/f/; /w/→/w/; /j/→/j/; /z/→/z/; /ʒ/→/z/.

### 2.3 Vocabulary

The development of language reflects the development progress of civilization in a society which can be seen in the development of vocabulary. For the Indonesian language, the development of vocabulary increased rapidly at the end of the 20th century and the beginning of the 21st century which, among other things, was driven by the development

of science, technology and the arts. This can be seen from the increase of entries in the official “Great Dictionary of the Indonesian Language” (Indonesian: KBBI) from one edition to the next, and for over the past 20 years, entries in the KBBI increased from 62,000 entries in the first edition (1988) to 91,000 entries in the fourth edition (2008) (Sugono, 2008).

## 3. Automatic Speech Recognition for Bahasa Indonesia

### 3.1 Indonesian ASRs and Their Phoneme Set

The first Indonesian ASR was a word-based system covering only a small vocabulary that was developed in 2004 for hearing and speaking impaired people (Sakti et al., 2004). It was performed on acoustic model based on the hidden Markov model with Gaussian mixture model (GMM-HMM). Then, a year later, the initial Indonesian phoneme-based ASR was designed using the cross-language approach, where English is the source language, and Indonesian is the target language (Sakti et al., 2005).

Since there were no agreed standard phoneme set for Indonesian ASR, many researches utilized their own phoneme set. Lestari et al. (2006) utilized 31 phonemes for their GMM-HMM based large-vocabulary ASR. The phoneme set was similar to (Soderberg, 2008) albeit with two differences. Phonemes /e/ and /ə/ were merged because they were spelled as ⟨e⟩. Phoneme /ʔ/ was not used while phoneme /q/ was introduced to accommodate the pronunciation of Arabic loanwords. Sakti et al. (2008a) also developed Indonesian large-vocabulary corpora and a large vocabulary ASR system using 32 phonemes similar to (Soderberg, 2008). They had been carried out for the A-STAR (Asian Speech Translation Advanced Research) consortium; the complete A-STAR system was successfully launched in 2010 (Sakti et al., 2008b; 2013).

Hoesen and Lestari (2014) did not employ the /q/ (since its number was too low and Indonesian tends to pronounce it as /k/) and the diphthongs. Clynes (1997) argued that Bahasa Indonesia (part of the Austronesian language family) did not have diphthongs. The diphthongs were thought of as a combination of a vowel and an approximant. Afterwards, Hoesen et al. (2016a; 2016b) performed experiment on various adaptation methods for the GMM-HMM model. The /q/ and the diphthongs were supplied back; phonemes /e/ and /ə/ were also treated as two separate phonemes. The reintroduction of the diphthongs was to avoid listing monophthongized version of every diphthong.

More recent research started to employ neural-network based acoustic model, the current state-of-the-art method. Hoesen et al. (2015; 2018) utilized shared-hidden-layer fully-connected neural-network (SHL-DNN) which was jointly trained with the high-resource English and the low-resource Indonesian data. The joint training could decrease the recognition errors for the low-resource Indonesian ASR. Although this research used not only Indonesian data, but also English data, phoneme set utilized in these researches were identical with (Hoesen and Lestari, 2014).

### 3.2 Dictionary and Language Model

Phonetic-based ASR system necessitates the use of phonetic dictionary. Phonetic dictionary translates a word or term into its phonetic sequence(s). Similar to the phoneme set, there is still no standard dictionary (including vocabulary) for Indonesian ASRs. Lestari et al. (2006; 2010) collected unique words from Indonesian news corpus that occurred more than 3 times. Meanwhile, Hoesen et al. (2016a; 2016b) collected unique words from KBBI, various reputable Indonesian news websites, and transcripts from various Indonesian speech corpora.

Some approaches were then performed to generate pronunciation to the words. Hoesen and Lestari (2014) and Hoesen et al. (2018) manually annotated each word in their dictionary. While this approach could yield the best phonetic accuracy, it was impractical for larger dictionary because it needed a considerable amount of time and resources.

As mentioned in Section 2, native Indonesian words tend to have a high degree of consistency between their spelling and pronunciation. Exploiting this tendency, Zahra et al. (2009) and Putri et al. (2019) specifically tried to automatically generate pronunciation for an Indonesian ASR dictionary using a set of rules. Their research yielded 3.2% and 12.71% phone error rates (PERs) respectively. The higher PER from (Putri et al., 2019) was caused by the occurrence of abbreviations and English words and loanwords, which could not be tackled thoroughly by the rules. Hoesen et al. (2019) tried to overcome these problems by employing seq2seq-based neural network to generate the pronunciation. Their cross-validation experiment could achieve 4.15-6.24% PER. To achieve better phone accuracy while requiring less time, automatic pronunciation generation then manual refinement can be performed, such as in (Hoesen et al., 2016a; 2016b).

Other than the dictionary, the majority of current ASRs require a language model (LM). In Indonesian, an LM for informal speech will be dissimilar from an LM for standard formal speech. Informal speech in Indonesian contains some disfluencies, a different set of vocabulary from the formal one, and unusual sentence structures (Hoesen et al., 2016a; 2016b). For example, the suffixes *-i* and *-kan* in formal speech becomes *-in* in informal speech. Complicating matters, Indonesian informal speech is also disparate from the informal text. Indonesian informal texts contain a considerable amount of unpronounceable abbreviations; thus, most informal texts cannot be used for training the spontaneous LM. One method to produce an LM more suitable for informal speech is to adapt a formal speech LM with spontaneous speech transcript (Lestari and Irfani 2015).

## 4. Automatic Speech Synthesizer for Bahasa Indonesia

Indonesian speech synthesizer (TTS, text-to-speech) has been developed for the past years. The early version of Indonesian TTS systems was developed using di-phone unit concatenation (Arman et al., 2001). Then, Sakti et al. (2008c) developed the first HMM-based speech synthesis for the Indonesian language using only limited resources. The subjective assessment of the Indonesian TTS in terms of both quality and intelligibility aspects had also been

conducted online by a web-based listening test system (Sakti et al., 2010).

After that, Jangtjik (2014) developed HMM-based speech synthesizer by adding English lexicon to deal with code-switching, using phone mapping technique i.e mapping English phonemes to their Indonesian equivalent as done by Lestari (2010). Hidayat (2015) improved the naturalness of the TTS by adding prosody model to the system. Improvement in acoustic model was also performed by Fanani (2017), employing DNN to improve overall naturalness. Gisela et. al. (2019) further developed the DNN TTS into an Indonesian-English polyglot TTS to ascertain code-switching problems. Research in Gisela et. al (2019) found that although it improves the intelligibility of the English words, in some cases, the non-polyglot, phone-mapping version of the English words were more easily understood.

### 4.1 Phoneme Set for Indonesian TTS

The phoneme set used for Indonesian TTS is generally the same, following the set defined by Soderberg (2008), although the vowels are quite simplified compared to the real-life case. The vowels /i, e, o, u/ have allophones /i, e, ə, u/ and they might depend on the regional accent of the speaker. Thus, usually in the TTS lexicon, the vowels /i, e, o, u/ represents both themselves and their allophones. The drawback of not including the allophone is that the generated sentences do not sound as natural as native speaker, especially since the allophones are often used in a final closed syllable, for example in the words 'sindir' /sindir/ and 'lumpur' /lumpuɾ/, or are used sporadically as in 'foto' /foto/ and 'tokoh' /təkoh/.

### 4.2 Vocabulary and Pronunciation Rules for Indonesian TTS

The vocabulary used for Indonesian TTS is quite similar with the ASR lexicon. However, since the TTS models only one speaker, there is no necessity to list multiple pronunciations for a single word. Pronunciations are generally taken from the KBBI. For words that are not provided by the KBBI, letter-to-sound rules are applied, since Indonesian has a quite straightforward grapheme-to-phoneme relationship. Some TTS even rely almost fully on letter-to-sound conversion algorithm, as done by Gisela et. al. (2019), except for words involving the letter 'e' and letter pairs 'ai', 'au', and 'oi' that are provided in a small lexicon.

## 5. Conclusion and Future Works

We have presented a summary of research activities that had been done on the Indonesian language, specifically on the design of the Indonesian phoneme set and two main speech technology applications (ASR and TTS). Nowadays, recent progress in Indonesia's research activities do not focus solely on Indonesian as the official language anymore but start to cover on various Indonesian ethnic languages spoken across the archipelago (i.e., Javanese, Sundanese, Balinese, and Batak languages).

## 6. Acknowledgements

This report is partially funded by the Ministry of Research and Higher Education of Indonesia under research project

with the title "Intelligent System to Monitor Gadget Usage in Teenagers using Machine Learning Technique". A part of the work on Indonesian ASR/TTS for A-STAR project had 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).

## 7. Bibliographical References

Arman, A. (2001). Prosody model for Indonesian text to speech system. Asia Pacific Conference on Communication, Tokyo.

Clynes, A. (1997). On the Proto-Austronesian "Diphthongs". *Oceanic Linguistics*, vol. 36(2), pages 347-362.

Hartanto R., Lestari D.P. (2019). Rule-based Approach for English-Indonesian Code-switching Acoustic Model. International Conference on Data and Software Engineering 2019.

Hasan, Alwi. Dendy, Sugono. Anton, Moeliono (1999). *Telaah Bahasa dan Sastra*. Yayasan Obor Indonesia, page 260.

Hoesen, D. and Lestari D.P. (2014). A Prototype of Indonesian Dictation Component for Typing and Formatting Document Using a Word Processor Software, Proceedings of International Conference on Electrical Engineering and Computer Science, pages 17-21.

Hoesen, D., Satriawan, C.H., Lestari, D.P., and Khodra, M.L. (2016a). Towards Robust Indonesian Speech Recognition with Spontaneous-Speech Adapted Acoustic Models. *Procedia Computer Science*, vol. 81, pages 167-173.

Hoesen, D., Lestari, D.P., and Khodra, M.L. (2016b). Adaptation of Acoustic Model for Indonesian Using Varying Ratios of Spontaneous Speech Data. Proceedings of the 2016 O-COCOSDA, pages 39-44.

Hoesen D., Price R., Lestari D.P., Shinoda K. (2015) A DNN-based ASR system for the Indonesian language. Proceedings of ASJ Autumn Meeting, pages 5-6.

Hoesen D., Lestari D.P., and Widyantoro D.H. (2018) Shared-hidden-layer deep neural network for under-resourced language. *Telkomnika*, vol. 16(3), pages 1226-1238.

Hoesen D., Putri, F.Y., and Lestari D.P. (2019) Automatic Pronunciation Generator for Indonesian Speech Recognition System Based on Sequence-to-Sequence Model. Proceedings of the 2019 O-COCOSDA.

Lestari, D.P., Iwano K., and Furui S. (2006). A large vocabulary continuous speech recognition system for Indonesian language. Proceedings of the 15<sup>th</sup> Indonesian Scientific Conference in Japan.

Lestari, D.P., Furui S. (2010). Adaptation to Pronunciation Variations in Indonesian Spoken Query-Based Information Retrieval, *IEICE Transactions on Information and Systems*, 93(9), pages 2388-2396.

Lestari, D.P. and Irfani A. (2015). Acoustic and language models adaptation for Indonesian spontaneous speech recognition. Proceedings of the 2nd International

Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA).

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (2013). *Ethnologue: Languages of the World*, Seventeenth edition. Dallas, Texas: SIL International.

Pastika, I. Wayan (2012). Pengaruh Bahasa Asing terhadap Bahasa Indonesia dan Bahasa Daerah: Peluang atau Ancaman?. *Jurnal Kajian Bali*, Vol. 02, No. 02, ISSN: 2580-0698.

Panitia Pengembang Pedoman Bahasa Indonesia (2016). Kementerian Pendidikan dan Kebudayaan. Pedoman Umum Ejaan Bahasa Indonesia. 4th ed.

Putri, F.Y., Hoesen D., and Lestari D.P. (2019). Rule-Based Pronunciation Models to Handle OOV Words for Indonesian Automatic Speech Recognition System. Proceedings of the 6th International Conference on Science in Information Technology (ICSITech).

Yap M.J. et al., (2010). The Malay lexicon project: A database of lexical statistics for 9,592 words., *Behavior Research Methods*, 42(4), pages 992-1003.

Sakti, S., Hutagaol, P., Arman, A.A., and Nakamura, S. (2004). Indonesian speech recognition for hearing and speaking-impaired people. International Conference on Spoken Language Processing. pages 1037-1040.

Sakti, S., Markov, K., and Nakamura, S. (2005). Rapid Development of Initial Indonesian Phoneme-based Speech Recognition Using the Cross-Language Approach. O-COCOSDA.

Sakti, S., Kelana, E., Riza, H., Sakai, S., Markov, K., Nakamura, S. (2008a). Recent Progress in Developing Indonesian Large-Vocabulary Corpora and LVCSR System. MALINDO. pages 40-45.

Sakti, S., Kelana, E., Riza, H., Sakai, S., Markov, K., Nakamura, S. (2008b). Development of Indonesian Large Vocabulary Continuous Speech Recognition System within A-STAR Project., *IJCNLP Workshop on TCAST*, pages 19-24.

Sakti, S., Maia, R., Sakai, S., Shimizu, T., Nakamura, S. (2008c). Development of HMM-based Indonesian Speech Synthesis. O-COCOSDA. pages 215-220.

Sakti, S., Sakai, S., Isotani, R., Kawai, H., Nakamura, S. (2010). Quality and Intelligibility Assessment of Indonesian HMM-Based Speech Synthesis System. MALINDO, pages 51-57.

Sakti, S., Paul, M., Finch, A., Sakai, S., Vu, T.-T., Kimura, N., Hori, C., Sumita, E., Nakamura, S., Park, J., Wutiwiwatchai, C., Xu, B., Riza, H., Arora, K., Luong, C.-M., Li, H. (2013). A-STAR: Toward Translating Asian Spoken Languages. Special issue on Speech-to-Speech Translation, *Computer Speech and Language Journal* (Elsevier), vol. 27, Issue 2, pages 509-527.

Soderberg C.D. & Olson K.S., (2008). Indonesian, *Journal of the International Phonetic Association*, 38(2), pages 209-213.

Sugono, Dendy., et al., (2008). *Kamus Pusat Bahasa*. Jakarta. ISBN 978-979-689-779-1

Zahra, A., Baskoro, S., and Adriani M. (2009). Building a pronunciation dictionary for Indonesian speech recognition system. Proceedings of Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST).