Development of HMM-based Indonesian Speech Synthesis

Sakriani Sakti, Ranniery Maia, Shinsuke Sakai,
Tohru Shimizu, Satoshi Nakamura

NICT / ATR Spoken Language Communication Research Laboratories,
2-2-2 Hikaridai, "Keihanna Science City", Kyoto 619-0288, Japan
Phone: +81 774 95 1345, Fax: +81 774 95 1308, URL: www.slt.atr.jp/slc-e/
{sakriani.sakti,ranniery.maia,shinsuke.sakai}@nict.go.jp
{tohru.shimizu,satoshi.nakamura}@nict.go.jp

Abstract

This paper presents a systematic study on the development of speech synthesis for the Indonesian language based on the hidden Markov model. The context-dependent speech HMMs are trained with two hours of phonetically-balanced utterances, which have been phonetically segmented by Viterbi alignment using an Indonesian speech recognition system. The contextual factors utilized here are mainly related to phoneme identity and its positional information with regard to word and sentence. This study did not use syllable and stress information. Although the quality of generated speech is "vocoded" buzzy speech, it is smooth and perfectly intelligible. Currently, this Indonesian speech synthesis has been incorporated in the mobile terminal of the NICT/ATR speech translation system.

Keywords: *hidden Markov model, Indonesian language, Speech synthesis.*

1. Introduction

Over the last decade, the most common speech synthesis technique used is based on a waveform concatenation algorithm, in which appropriate subword units are selected from speech databases [1]. This technique has been shown to synthesize high quality speech and is used for many applications. The Indonesian speech synthesis systems reported so far have also been developed using diphone unit concatenation [2].

Although the unit concatenation technique has met with significant success, it still holds certain limitations. The prosody and expressiveness of the speech output are limited to the original contents of the underlying database, and the synthesizer cannot generally extrapolate the instances of a speech unit, which were not visible during training. Thus, it is also not possible to guarantee that a discontinuity of inappropriate units may not occur in synthesized speech output. Consequently, in order to synthesize speech with various characteristics (e.g., speaking styles or emotions), as well as reach a high quality of speech itself, a large amount of speech data is required to cover all instances of desired prosodic, phonetic and stylistic variation. However, such resources require considerable time and efforts, which are difficult to obtain.

Recently, a statistical parametric speech synthesis system was proposed; this system, which has gained popularity, is based on the hidden Markov models (HMMs) in which speech waveforms are generated through parameters directly obtained from the HMMs [3]. This system offers the ability to model different speech styles without the need for recording very large databases. It can be carried out by appropriately transforming the HMM parameters, using either speaker adaptation or interpolation techniques [4, 5]. Furthermore, although it has been originally developed to support the Japanese language, this system has been successfully applied to various languages such as English [6], Portuguese [7], Thai [8], etc.

This paper presents our systematic study of the development of HMM-based speech synthesis for the Indonesian language. In the next section, we briefly describe the characteristics of the Indonesian language, including the Indonesian phonological system. Issues pertaining to data resources, such as database design of phonetically-balanced sentences and speech recording process are described in Section 3. The development of an HMM-based speech synthesis system is presented in Section 4. Then, the integration of this speech synthesis system in a handheld speech translation system is described in Section 5. Finally, we draw our conclusions in Section 6.

2. Characteristic of Indonesian Language and Phonological System

The Indonesian language, so-called Bahasa Indonesia, is a unity language formed from hundreds of languages spoken in the Indonesian archipelago. Compared to other languages, which have a high density of native speakers,

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		

Indonesian is spoken as a mother tongue by only 7% of the population; more than 195 million people speak Indonesian as a second language with varying degrees of proficiency. Approximately, there are 300 ethnic groups living in 17,508 islands, that speak 365 native languages and no less than 669 dialects [9].

The language structure of Bahasa Indonesia is fairly simple when compared to that of some other languages. Unlike the Chinese language, it is not a tonal language. It is a language with neither declensions nor conjugations. It uses the same subject-verb-object word order used in English. Nouns have no gender and do not require any article. A plural noun is simply expressed by means of reduplication. Adjectives always follow the noun, while verbs are not inflected for person or number. Further, there are no tenses; tense is denoted by time adverbs or by other tense indicators, such as "*sudah*" (meaning "*already*") or "*belum*" (meaning "*not yet*"). The easiest way to make a question is to merely add a question mark and use a rising intonation [10].

Bahasa Indonesia is phonetic based and written in Roman script, which uses 26 letters similar to the English/Dutch alphabet. All letters are pronounced much more consistently, and no letters are muted. A peculiarity in the spelling of this language is the lack of a separate sign to denote the schwa. Both phonemes /e/ and the schwa /ə/ are written as an "e," which can occasionally be confusing.

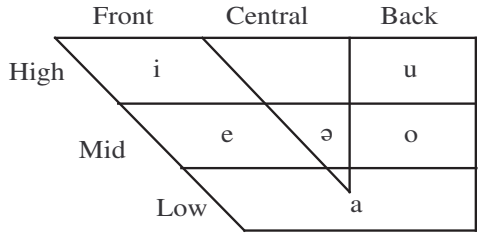

Figure 1: Articulatory pattern of Indonesian vowels.

The full phoneme set, as defined in an Indonesian grammar text [11], contains a total of 33 phoneme symbols. They consist of 10 vowels (including diphthongs), 22 consonants and one silence symbol. The articulatory pattern of Indonesian consonants is given in Table 1, and Fig. 1 illustrates the vowel articulation pattern. The vowel pattern indicates the first two resonances of the vocal tract, F1 (height) and F2 (backness), which consist of /a/ (like "a" in "father"), /i/ (like "ee" in "screen"), /u/ (like "oo" in "soon"), /e/ (like "e" in "bed"), /ə/ (a schwa sound, like "e" in "learn"), /o/ (like "o" in "boss") and four diphthongs, /ay/, /aw/, /oy/ and /ey/.

Indonesian word stress typically falls on the pre-final syllable, unless this syllable contains a schwa in which case, the stress is final. However, free variation of stress position is commonly observed, since speakers with different ethnic native languages may behave differently with respect to stress realization and perception [12]. Fortunately, unlike the case in many Western languages, the word stress in Indonesian is phonetically weakly marked. No phonological rules, structural or contrastive differences based on stress are observed. Similarly, there are no words containing the same sequence of vowels and consonants that differ in their stress patterns and, consequently, in their meanings. The difference in duration between stressed and unstressed syllables is also comparatively small. Experiments by some researchers [13] indicated that Indonesian listeners are relatively tolerant with regard to stress and its position. They even concluded that the stress might be communicatively irrelevant or essentially free in Indonesian.

3. Data Resources

3.1. Text Corpus

Two types of text data are used here, including:

1. **Travel expression task**

   The ATR basic travel expression corpus (BTEC) has served as the primary source for developing broad-coverage speech translation sys-

tems [14]. The sentences were collected from Japanese/English sentence pairs in travel domain "phrasebooks" by bilingual travel experts and have been translated into several languages including French, German, Italian, Chinese, Korean and Indonesian. For this speech synthesis development, 510 sentences of Indonesian BTEC1 were selected.

2. **Daily news task**

   A raw text source for the daily news task has already been generated by an Indonesian student [15]. The source was compiled from "KOMPAS" and "TEMPO," which are currently the largest and most widely used Indonesian newspaper and magazine, respectively. The raw text source consisted of more than 3,160 articles with about 600,000 sentences.

3.2. Speech Corpus

We first selected phonetically-balanced sentences from both the text data described above, which assumed to cover almost all of phonetic contexts used in the Indonesian language. Using the greedy search algorithm [16], a total of 2,012 sentences are produced. The number of units and coverage rate of the training data that obtained in the resulting sentences are shown in Table 2. After that, we recorded these sentences, uttering by a female Indonesian speaker who spoke standard Indonesian (no accent). The speech recording was conducted in a sound proof room, at a 48 kHz sampling rate with 16 bits resolution. The sampling rate was later downsampled to 16 kHz for our experiments.

Table 2: *Number of units and coverage rate of the training data that obtained in the resulting 2,012 sentences.*

Phone	# Units	Coverage
Monophones	33	100%
Left Biphones	814	99.75%
Right Biphones	813	99.75%
Triphones	8270	85.18%

4. Speech Synthesis Experiments

These experiments were conducted using an open source speech synthesis engine, known as HMM-based Speech Synthesis System (HTS) [17]. The complete process consists of two parts: training and synthesis, which are illustrated in Figure 2. Both parts are briefly explained in the following sections.

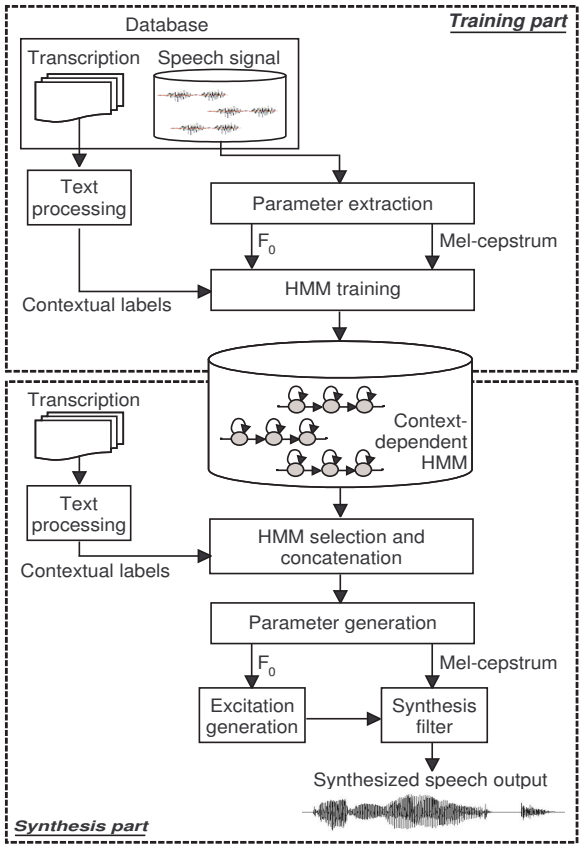

Figure 2: An HMM-based speech synthesis system, which consists of a training and a synthesis parts.

4.1. The Training

The models were trained using the two hours of speech material described in Section 3, and the training consisted of following processes:

1. **Utterances Segmentation**

   As is the case with speech recognition systems, segmented utterances according to phonetic labels are generally used as a starting point for training speech models. In this study, this was automatically done by Viterbi alignment of the spoken utterances and the corresponding transcription using the Indonesian speech recognition system [18].

2. **Parameter Extraction**

   The speech signals were sampled at a rate of 16 kHz and windowed by a 25-ms Hamming window. Then, both excitation and spectrum parameters are

extracted from the speech database at a frame of every 5-ms. The excitation feature vector (pitch) consisted of log F_0 and its dynamic parameters (delta and acceleration). The spectral feature vector consisted of 25 mel-cepstral coefficients [19], including the zeroth coefficient, and their dynamic parameters (delta and acceleration).

3. **Contextual Labels Generation**

There are many contextual factors (e.g., phoneme identity, word stress, etc.) that might have an effect on the prosodic characteristic of speech. In this preliminary study, it is only the phoneme identity and its positional information with regard to word and sentence that are taken into account. This study did not use syllable and stress information, since stress is phonetically weakly marked and can be considerably free in Indonesian (see Section 2). Part of speech (POS) tagging was also not yet included. The full contextual label was generated from a phonetic transcription using text processing tools, and it consisted of the following contextual factors:

- Phoneme level:
  - {second preceding, preceding, current, succeeding, second succeeding} phoneme;
  - position of current phoneme in the current word (forward and backward);
- Word level:
  - number of phonemes in {preceding, current, succeeding} word;
  - position of current word in the current utterance (forward and backward);
- Utterance level:
  - number of words in the utterance;
  - utterance types: declarative, interrogative or imperative sentence.

4. **Context-dependent HMM Modeling**

Five state left-to-right HMMs were used, where each HMM corresponds to a phoneme-sized speech unit. These context-dependent HMMs were trained using the full contextual labels and the concatenated feature vectors of extracted F_0 and mel-cepstrum parameters. The mel-cepstrum feature vectors were modeled by continuous probability distribution, while the F_0 feature vector were modeled by multi-spaced probability distribution (including a discrete voiced/unvoiced symbol and one-dimensional continuous log F_0 values). The state durations of each HMM were modeled by n-dimensional Gaussians, where the dimension was equal to the number of states of the HMM.

5. **Decision-tree Context Clustering**

Since there is a wide range of different contextual labels, it is obvious that the amount of training data is insufficient for obtaining a reliable estimate for all model parameters. In a manner similar to speech recognition, the clustering technique may be utilized to overcome this problem. Here, the distributions for the excitation (pitch) parameter, spectral parameter and the state duration were clustered independently using a decision-tree based context clustering technique. By applying 1250 phonetic and positional questions, the resulting trees for spectrum, pitch and duration models had 2,409 leaves, 4,245 leaves and 961 leaves, respectively.

4.2. The Synthesis

With regard to the synthesis, the system first converted a given arbitrary input sentence into a contextual label sequence. Then, three sets of context-dependent HMMs, for each of the F_0, mel-cepstrum and duration parameters, were selected and concatenated according to the label sequence. Parameter durations were determined in such a way that the HMM output probability was maximized. Following this, F_0 and mel-cepstrum were generated, based on the obtained state durations. Finally, a speech waveform was directly synthesized from the obtained F_0 values and mel-cepstral coefficients by using a synthesis filter.

Fig. 3 shows an example of the spectrogram comparisons of both natural speech (top) and synthesized speech (bottom) for an utterance "*Saya berencana untuk pergi ke konser malam ini*" (meaning "*I plan on going to the concert this evening*") which is part of the training data. It is observed that the system is able to synthesize speech that resembles the speaker's speech in the database. The speaking rate of the synthesized version is also similar to that of the natural speech case. Through informal listening tests, we have found that the quality of generated speech is typical "vocoded" buzzy speech. The lack for a separate sign to denote the phonemes /e/ and the schwa /ə/ also affect the synthesized speech output. However, by and large, the sound is smooth and perfectly intelligible.

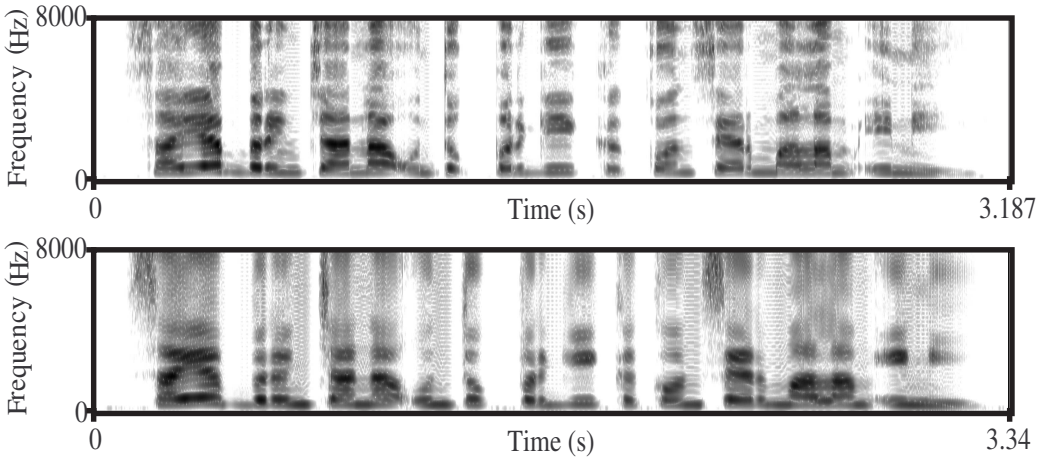

Figure 3: Spectrograms of both natural speech (top) and synthesized speech (bottom) for an utterance "*Saya berencana untuk pergi ke konser malam ini*" (meaning "*I plan on going to the concert this evening*").

5. Integration in a Hand-held Speech Translation System

The obtained Indonesian HMM-based speech synthesis system has currently been integrated in the mobile terminal of the NICT/ATR Indonesian-Japanese speech translation system. It is designed for a practical use as a translation assistance tool for traveling abroad. Fig. 4 shows the entire speech translation system, which consists of several components including Japanese speech recognition system, Indonesian speech recognition system, Japanese-Indonesian machine translation system, Japanese speech synthesis system and Indonesian speech synthesis system. This system will translate utterance by utterance in a real-world environment. The input speech of source language is recognized using the speech recognizer. Then, the resulting text is translated into a target language by the machine translator. Finally, the synthesizer is used to produce the spoken output.

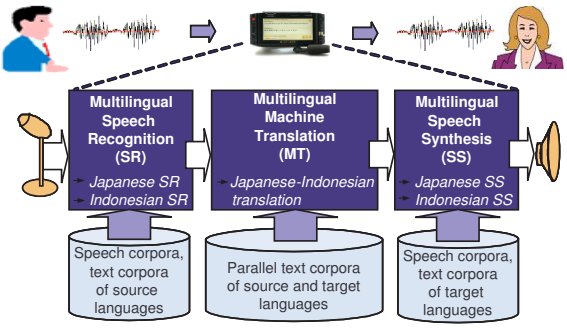

Figure 4: An architecture of NICT/ATR Indonesian-Japanese speech translation system, including speech recognition, machine translation and speech synthesis.

This project was carried out in accordance with the "Asia speech translation (A-STAR) Consortium" [20]. The goal of the project is to advance the development of multilingual man-machine interfaces, particularly of multilingual speech translation systems, in the Asian region. Thus, the final speech translation system is expected to include not only Japanese and Indonesian languages, but also many other languages from Asian countries. These fundamental technologies are expected to be applicable to the human-machine interfaces of various telecommunication devices and services connecting many Asian countries through a network. The improvements in borderless communication in the Asian region are expected to have benefits in many fields including tourism, business, education and social security.

6. Conclusion

In this paper, we have presented the development of an Indonesian HMM-based speech synthesis system. The system was trained with only limited speech data and a few contextual factors; however, it is able to synthesize speech which resembles the speaker's in the training database. The speaking rate of the synthesized version is also similar to the natural speech case. Although the quality of generated speech is "vocoded" buzzy speech, it is smooth and perfectly intelligible. The entire process of development took only one month, which proves that the HMM approach is very effective in the rapid development of the speech synthesis system for new languages. The system was also successfully integrated into a hand-held speech translation system.

Future work in this field involves the utilization of a larger speech corpus with wider contextual factors such as syllables, stress, phrases and POS tags, as well as the investigation of the values of those information to

the synthesizer . Experiments related to the synthesis of voices with different accents of ethnic native languages might also prove to be useful. The formal evaluation of intelligibility, naturalness and functionality of synthesized speech by human experts and users will also be carried out in order to obtain the rate of the overall quality.

7. References

[1] N. Campbell and A. Black, "Prosody and the selection of source units for concatenative synthesis," in *Progress in speech synthesis*, J. van Santen, R. Sproat, J. Olive, and J. Hirschberg, Eds. 1996, pp. 279–282, Springer Verlag.

[2] A. A. Arman, "Prosody model for indonesian text to speech system," in *Proc. Asia Pacific Conference on Communication*, Tokyo, Japan, 2001.

[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, Istanbul, Turkey, 2000, pp. 1315–1318.

[4] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *Proc. ICASSP*, Salt Lake City, Utah, USA, 2001, pp. 805–808.

[5] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura, "Speaker interpolation for HMM-based speech synthesis system," *J. Acoust. Soc. Jpn. (E)*, vol. 21, no. 4, pp. 199–206, 2000.

[6] K. Tokuda, H. Zen, and A.W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Speech Synthesis Workshop*, Santa Monica, California, USA, 2002.

[7] R. Maia, H. Zen, K. Tokuda, T. Kitamura, and F. G. V. Resende Jr., "An HMM-based Brazilian Portuguese speech synthesizer and its characteristics," *IEEE Journal of Communication and Information Systems*, vol. 21, no. 2, pp. 58–71, 2006.

[8] S. Chomplan and Takao Kobayashi, "Implementation and evaluation of an HMM-based Thai speech synthesis system," in *Proc. EUROSPEECH*, Antwerp, Belgium, 2007, pp. 2849–2852.

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

[10] S. Backshall, *Indonesia*, Rough Guides, 2003.

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

[12] R. Goedemans and E. van Zanten, "Stress and accent in indonesian," in *Malay / Indonesian Linguistics*, D. Gil, Ed., London, UK, 2007, Curzon Press.

[13] E. van Zanten and V. J. van Heuven, "Word stress in Indonesian; its communicative relevance," *Journal of Humanities and Social Sciences of Southeast Asia and Oceania*, , no. 154, pp. 129–147, 1998.

[14] 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.

[15] 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.

[16] 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.

[17] "The HMM-based speech synthesis system (HTS)," http://hts.ics.nitech.ac.jp.

[18] S. Sakti, E. Kelana, H. Riza, S. Sakai, K. Markov, and S. Nakamura, "Development of indonesian large vocabulary continuous speech recognition system within A-STAR project," in *Proc. Workshop on Technologies and Corpora for Asia-Pacific Speech Translation*, Hyderabad, India, 2008, pp. 19–24.

[19] K. Tokuda, T. Kobayashi, and S. Imai, "Adaptive cepstral analysis of speech," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 6, pp. 481–489, 1995.

[20] 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, pp. 45–46.