Quality and Intelligibility Assessment of Indonesian HMM-Based Speech Synthesis System

Sakriani Sakti, Shinsuke Sakai, Ryosuke Isotani, Hisashi Kawai, Satoshi Nakamura
NICT Spoken Language Communication Research Group
3-5 Hikaridai, "Keihanna Science City", Kyoto 619-0289, Japan
{sakriani.sakti,shinsuke.sakai,ryosuke.isotani,hisashi.kawai,satoshi.nakamura}@nict.go.jp

Abstract

Recently, we have reported the pioneering development of an Indonesian speech synthesis system based on the hidden Markov models (HMMs). Through informal listening tests, we have found that the prosody was considerably good and the synthesized speech sounded smooth and stable. However, it is still necessary to assess whether the performance of the system is sufficient for applications where users may not be assumed to have any prior exposure to synthesized speech. In this paper, we present the formal assessment of Indonesian speech synthesis system in terms of both quality and intelligibility aspects. The assessment was done specifically using the mean opinion score (MOS) and semantically unpredictable sentence (SUS) tests, online by a web-based listening test system. Fifteen Indonesian subjects were involved. These users had no prior training and most of them were not familiar with speech synthesizers. The results showed that even for the smallest system which was trained only with 12 minutes of speech, the speech quality reached an MOS level of 2.78 and a word accuracy level of 90.48% in the SUS test.

1. Introduction

Over the last decade, the most commonly used speech synthesis technique is based on a waveform concatenation algorithm, in which appropriate subword units are selected from speech databases [1]. Some works on Indonesian speech synthesis systems based on diphone unit concatenation also exist [2]. The major advantage of this technique is the capability to synthesize high quality speech. However, to synthesize speech with various characteristics, as well as reach a high quality of speech itself, a large amount of speech data is required. Thus it often faces implementation obstacles when applying on some platforms having limitations of computational cost or memory footprint.

Tokuda et al. [3] have proposed a statistical parametric speech synthesis system. This system which recently has gained popularity, is based on the hidden Markov models (HMMs) in which speech waveforms are generated through parameters directly obtained from the HMMs. The advantage is that the 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.

Recently, we also have reported the pioneering development of an Indonesian speech synthesis system based on the hidden Markov models (HMMs) [9]. Through informal listening tests we have found that the prosody was considerably good and the synthesized speech sounded smooth and stable. However, it is still necessary to assess whether the performance of the system is sufficient for applications where users may not be assumed to have any prior exposure to synthesized speech. In this paper we present the formal assessment of Indonesian speech synthesis system in terms of both quality and intelligibility aspects. The assessment was done specifically using the mean opinion score (MOS) and semantically unpredictable sentence (SUS) tests, online by a web-based listening test system.

The rest of this paper is organized as follows. Section 2 provides an overview of the characteristics of the Indonesian language. Section 3 describes issues pertaining to data resources, such as database design of phonetically-balanced sentences and the speech recording process. Section 4 describes the development of an HMM-based speech synthesis system. Section 5 provides an evaluation of the generated speech using subjective listening tests. Finally, Section 6 presents the conclusion.

2. Characteristic of Indonesian Language

The Indonesian language, so-called Bahasa Indonesia, is a unity language formed from hundreds of languages spo-

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		

ken in the Indonesian archipelago. Compared to other languages, which have a high density of native speakers, 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. There are approximately 300 ethnic groups living in 17,508 islands that speak 365 native languages and no less than 669 dialects [10].

The language structure of Bahasa Indonesia is fairly simple in comparison to 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. 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 [11].

Bahasa Indonesia is phonetically based and written in Roman script with 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 phoneme *schwa*. Both phonemes /e/ and the *schwa* /ə/ are written as an "e," which can occasionally be confusing.

The full phoneme set, as defined in an Indonesian grammar text [12], 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/.

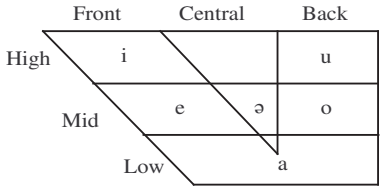

Figure 1. Articulatory pattern of Indonesian vowels.

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 [13]. Fortunately, unlike 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 [14] 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 systems [15]. 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 [16]. 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 the text data described above, which assumed to cover almost all phonetic contexts used in the Indonesian language. Using the greedy search algorithm [17], 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.

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%

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. Finally, the speech was organized into three different training sets of different sizes; sets with a duration of 12 minutes, 60 minutes, and 120 minutes.

4. Development of Speech Synthesis System

These experiments were conducted using an open source speech synthesis engine, known as HMM-based Speech Synthesis System (HTS) [18]. 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.

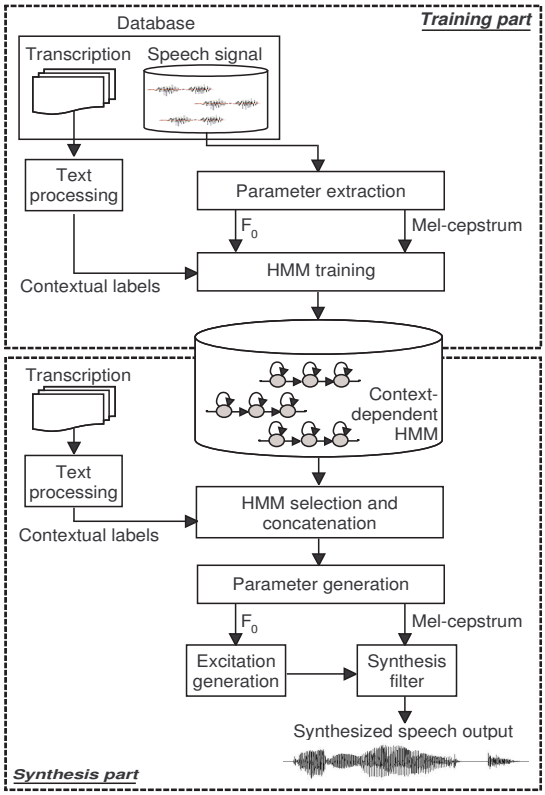

Figure 2. An HMM-based speech synthesis system, which consists of a training and a synthesis parts.

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 [19].

2. **Parameter Extraction**

   The speech signals were sampled at a rate of 16 kHz, windowed using a 25-ms Hamming window and 5-ms frame shift. The feature vector consisted of excitation (pitch) and spectral parameters. 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, including the zeroth coefficient, and their dynamic parameters (delta and acceleration). The mel-cepstral coefficients were obtained by mel-cepstral analysis [20]. Additionally, we investigated also the use of the smoothed spectrum analyzed by speech transformation and representation using adaptive interpolation of weighted spectrum (STRAIGHT) [21]. Table 3 shows all speech synthesis systems with different training sets and spectrum parameters.

Table 3. Various speech synthesis systems with different training set sizes and spectrum parameters.

SS system	Spectrum parameters	Training data (∼min of speech)
A	24 MCEP	12
B	24 MCEP	60
C	24 MCEP	120
D	24 STRAIGHT MCEP	120

3. **Contextual Label Generation**

   There are many contextual factors (e.g. phoneme identity, word stress, etc.) that might have an effect on the prosodic characteristic of speech. The contextual factors utilized here were mainly related to phoneme identity and phonemes' positional information with regard to word and sentence. This study did not involve the use of syllable, phrase and part of speech (POS) tagging information. Word stress information was also not included since, in Indonesian, stress is phonetically weakly marked and can be considerably free (see Section 2). Full contextual labels were generated from phonetic transcription and word boundary information using text processing tools; the labels had the following features:

   - 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 are many combinations of contextual features, the model parameters may not be reliable when they are estimated using only limited training data. 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. A total of 1250 questions were applied according to the following major distinctions:

   - **Questions based on articulatory phonetics described in Table 1 and Fig. 1**
     - *Is the current phoneme a plosive consonant?*
     - *Is the succeeding phoneme a middle vowel?*
   - **Questions based on positional factors**
     - *Is the current phoneme in position 1 within the current word?*
   - **Questions based on quantitative parameters**
     - *Is the number of phonemes in the previous word 6?*
     - *Is the number of words in the sentence 10?*
   - **Questions based on sentence type**
     - *Is the sentence a declarative sentence?*

The summarization of the resulting spectrum, pitch and duration decision-trees are described in Table 4.

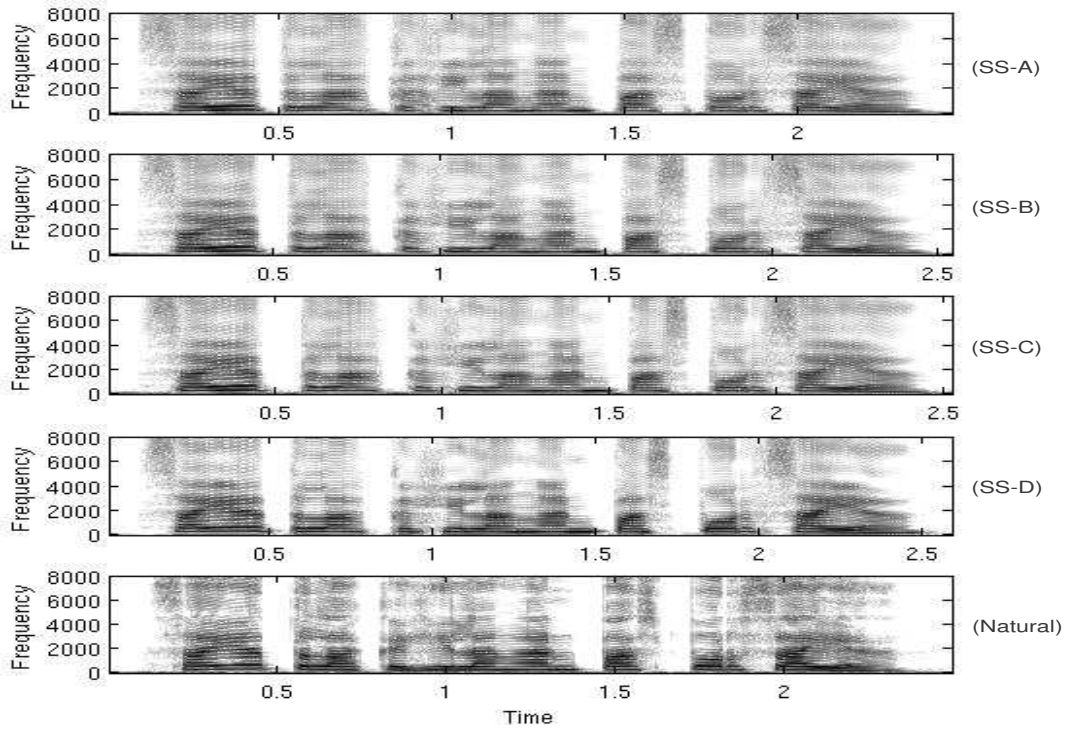

Figure 3. Spectrogram comparisons of natural speech and all speech synthesis systems (A, B, C, and D) for an utterance "*Saya telah kehilangan paspor saya*" (meaning "*I've lost my passport*").

Table 4. Number of leaves of the resulting spectrum, pitch and duration decision-trees.

SS system	# Tree leaves		
	Spectrum	Pitch	Duration
A	431	876	179
B	1558	2800	573
C	2485	4374	943
D	2116	4178	869

4.2. The Synthesis

The system is able to synthesize from arbitrary input sentences in the following way: first, convert the input sentence into a contextual label sequence using Indonesian text processing (a grapheme-to-phoneme conversion was also carried out here in order to deal with the out-of-vocabulary words); next, select and concatenate three sets of context-dependent HMMs for the F_0, mel-cepstrum, and duration parameters respectively, according to the label sequence; finally, synthesize a speech waveform directly from the obtained parameters by using only a simple excitation and the mel-log spectrum approximation (MLSA) filter [20].

Fig. 3 shows an example of the spectrogram comparisons of natural speech and all speech synthesis systems (A, B, C, and D) for a new utterance "*Saya telah kehilangan paspor saya*" (meaning "*I've lost my passport*") which has not been learnt during training. 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.

5. Assessment of Synthesized Speech

The assessment of synthesized speech was conducted online by a web-based listening test system. There were 15 Indonesian subjects ranging from 20 to 40 years. They had no prior training and most of them were not familiar with speech synthesizers. Two major aspects of the speech synthesis system evaluation are discussed in the following sections.

5.1. Overall Quality Assessment

Here, we employed the most commonly used method of evaluation, namely, the mean opinion score (MOS) [22] test. Both the travel expression and daily news text genres were subjected to the synthesis. Subjects listened to each presented speech and were required to rate the overall quality with regard to aspects such as acceptability, naturalness, and clarity. A 5-point MOS scale was used, where 5 indicated 'excellent' (the speech utterance sounds very clear and perfectly natural) and 1 indicated 'bad' (the speech utterance sounds unclear and completely unnatural). Each speech utterance could be played as many times as the subjects wished. There were two sessions involving the use of the MOS test. First, we evaluated Systems A, B, and C in order to investigate the speech quality of various synthesis systems which were trained on different training set sizes. Then, in the second session, we investigated the speech quality produced by the systems with different spectrum parameters in comparison with the quality of natural speech. Here, Systems C and D were evaluated. In total, in each session, there were 45 speech utterance (15 utterances per system), which were presented in random order.

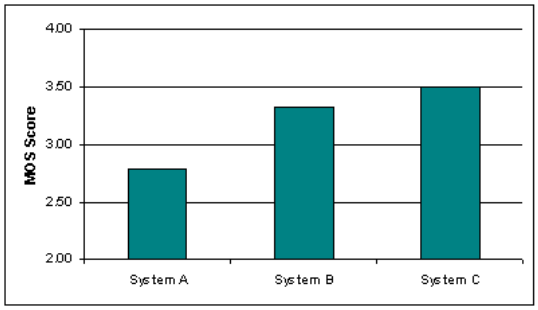

Figure 4. Overall MOS quality results for Systems A, B, and C.

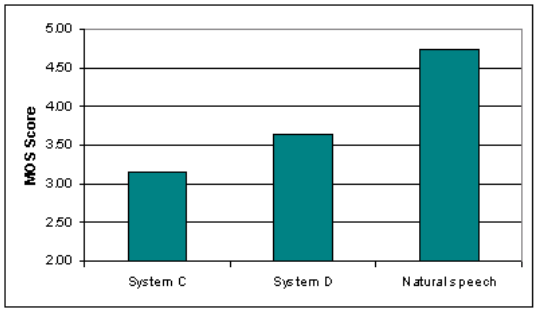

Figure 5. Overall MOS quality results for Systems C and D, and the natural speech.

The MOS results for Sessions 1 and 2 are presented in Fig. 4 and Fig. 5, respectively. Considering that the subject listeners were not familiar with speech synthesizers, Fig. 4 shows that the overall quality of all the systems was generally fairly good. Even for the smallest system, namely, System A, which was trained with only 12 minutes of speech, the speech quality reached to an MOS level of 2.78. When the training was conducted with more than 60 minutes of speech, the quality difference between Systems B and C was no longer large. However, there was still a significant difference in quality compared to the quality of natural speech, as shown in Fig. 5. Here, it is evident that applying STRAIGHT spectral analysis to System D enhanced the speech quality of the synthesis.

5.2. Intelligibility Assessment

For assessing the intelligibility of the speech produced by the systems, a semantically unpredictable sentences (SUS) [23] test was used, in which each sentence had a valid syntactic structure but was semantically nonsensical. A typical English SUS sentence has a 'det adj noun verb det adj noun' structure. Here, we slightly modified the structure to fit Indonesian grammar and the resultant structure was 'num-det noun adj verb num-det noun adj'. The words were randomly selected from an Indonesian word list. An example of an Indonesian SUS sentence is "*Setiap pagi hitam ditutup berbagai danau tua*" (meaning "*Each black morning was closed by various old lakes*"). In this case, each speech utterance could be played only once or twice at the most. Then, the subjects had to write down all the words in the sentence as best as they could.

Table 5. SUS intelligibility accuracy results for Systems A, B, C, and D.

SS system	# Errors			Sent Acc	Word Acc
	# Sub	# Ins	# Del		
A	36	0	14	54.67%	90.48%
B	28	0	3	66.67%	94.10%
C	16	0	5	78.67%	96.00%
D	19	0	5	78.67%	95.42%

The SUS test was conducted for Systems A, B, C, and D. In total, 80 speech utterances were synthesized, out of which 20 speech utterance (5 utterances from each system) were chosen and presented randomly to each listener. The results for each system are presented in Table 5. A sentence was considered to be transcribed correctly only if all the words were correctly transcribed. Therefore, it is not surprising that the overall intelligibility at the sentence level

was not very high. This may have been because it was quite challenging for the subjects to guess all the words correctly by listening to the speech only once or twice. The results revealed that most of the errors arosed because the subjects transcribed substitutions of various words instead of the actual words. Nevertheless, the intelligibility at the word level was rather high, and all systems reached a word accuracy level of above 90%. The intelligibility rate of System C was higher than that of System D, while the overall MOS quality rate of System D was higher than that of System C. This may reveal that the correlation between 'intelligibility' and 'naturalness' was not really strong; a natural sound quality may not always be intelligible. However, since the difference in word accuracy is not significant, we may still consider them as equally intelligible.

6. Conclusion

The subjective assessment of Indonesian speech synthesis system in terms of both quality and intelligibility aspect have been conducted using MOS and SUS tests, online by a web-based listening test system. The results showed that even for the smallest system, namely, system A which was trained with 12 minutes of speech and limited contextual factors, the speech quality reached an MOS level of 2.78 and a word accuracy level of 90.48% in the SUS test. It is revealed that this low resources Indonesian HMM-based speech synthesis systems were capable of producing highly intelligible natural Indonesian speech of good quality. The optimum performance was obtained by the system, using STRAIGHT spectral analysis. This system was successfully integrated into a hand-held NICT network-based speech translation system.

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] S. Sakti, R. Maia, S. Sakai, T. Shimizu, and S. Nakamura, "Development of HMM-based Indonesian speech synthesis," in *Proc. Oriental COCOSDA Workshop*, Kyoto, Japan, 2008, pp. 215–220.

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

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

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

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

[21] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné, "Restructuring speech representations using a pitchadaptive time-frequency smoothing and an instantaneousfrequency- based f0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[22] CCITT, *Absolute category rating (ACR) method for subjective testing of digital processors*, Red Book, 1984.

[23] C. Benoit and M. Grice, "The SUS test: a method for the assesment of text-to-speech intelligibility using semantically unpredictable sentences," *Speech Communication*, vol. 18, pp. 381–392, 1996.