Development of HMM-based Malay Text-to-Speech System

Zhi-Zheng Wu[1], Eng Siong Chng[1], Haizhou Li[2]
[1] School of Computer Engineering, Nanyang Technological University, Singapore
[2]Human Language Technology Department, Institute for Infocomm Research, Singapore
{wuzz, ASESChng}@ntu.edu.sg, hli@i2r.a-star.edu.sg

***Abstract*—This paper presents the development of a hidden Markov model (HMM)-based Malay text-to-speech (TTS) system. To our knowledge, this is the first report on the development of the HMM-based speech synthesis system for the Malay language. In this paper, We first discuss the Malay speech characteristics, specifically, on Malay phonological system and syllable structure. In the Malay phonological system, 37 phonemes are adopted as the phonemic representations. Then, we describe a HMM-based TTS framework and language specific knowledge such as phonological, linguistic information, and utterance structure, which is used in context dependent continuous HMM and tree-based clustering. After that, we report the development of Malay TTS corpora. Finally, a male and a female HMM-based Malay TTS systems are developed and evaluated. We further conduct listening test based on the Mean Opinion Score (MOS), and the results show that the developed HMM-based Malay TTS system can generate speech with acceptable quality in terms of naturalness and intelligibility.**

Index Terms: speech synthesis, statistical parametric speech synthesis, hidden Markov model, Malay Language.

I. Introduction

Speech synthesis is an important component for providing flexible human-computer interaction interface. To make communication between a machine and a person more natural and efficient, text-to-speech (TTS) synthesis systems are required to have an ability to generate natural sounding speech with arbitrary speaker's voice and synthesize speech of different languages.

Hidden Markov model (HMM)-based speech synthesis framework has been proposed by Tokuda's group[1,2], providing a simple and flexible framework to adapt to new languages and new speaking style. In this framework, the spectral envelop, fundamental frequency and duration are modeled simultaneously by HMMs[2]. During synthesis, given a text sequence, speech representation parameters are generated from the trained HMMs in the Maximum Likelihood (ML) sense. In the past years, the HMM-based speech synthesis approach has been improved by the speech community and has become increasingly popular. Many techniques which are developed for HMM-based speech recognition have been successfully applied to the HMM-based speech synthesis, such as context-dependent modeling[3], decision tree based clustering[4]; and issues such as vocoder $buzzy$ problem, acoustic modeling accuracy, over-smooth problem which adversely affect speech quality have been addressed in one way or another[4].

Currently, HMM-based speech synthesis framework has been successfully applied to different languages, such as English[5], Mandarin Chinese[6], Japanese[2], Thai[7], French[8], German[9], Arabic[10] and so on[4]. To apply the HMM-based speech synthesis framework to a new language, one of the main emphasis is to investigate the required language so as to efficiently model the contextual information as their contextual factors are different[4]. For example, tone contexts factors should be investigated for tonal languages[5]. In most of these cases, the developed text-to-speech system can perform well in terms of generating speech with acceptable quality. In this work, we describe the development of the HMM-based speech synthesis framework for the case of the Malay language, which is the national language of Malaysia and one of four official languages of Singapore.

In this paper, Malay speech characteristics are discussed first, and then HMM-based speech synthesis framework and language contextual factors are introduced in section III. At last, corpus collection, experimental setup and evaluation result are presented and a conclusion is given.

II. Malay Speech Characteristics

The Malay Language is a branch of Austronesian family of languages. It's a non-tonal language and without lexical stress. When developing TTS system, phonetic information is necessary. In this section, phonetic aspects, phonemic representations and syllable structure, of Malay Language to be used in the TTS system are discussed.

A. Malay Phonological System

TABLE I
Malay Vowel and Diphthong System

		Vowel Backness		
		Front	Central	Back
Vowel Height	High	i		u
	Medium	e	ə	o
	Low		a	
Diphthongs		ai,oi		au

A set of 37 phonemes are adopted as the phonemic representation of Malay phonemes. These phonemes are 6 vowels, 27 consonants, 3 diphthongs and one for silence. In Table I and II, the Malay vowels, consonants and diphthongs are illustrated in the International Phonetic Alphabet (IPA) and grouped by the articulation attributes. In Table I, vowels

Proceedings of the Second APSIPA Annual Summit and Conference, pages 494–497,
Biopolis, Singapore, 14-17 December 2010.

TABLE II
MALAY CONSONANT SYSTEM

Manner of articulation		Place of articulation							
		Labial		Coronal			Dorsal		
		Bilabial	Labiodental	Dental	Alveolar	Postalvedar	Palatal	Velar	Glottal
Stops	Voiceless	p			t			k	ʔ
	Voiced	b			d			g	
Non-stops	Nasale	m		n			ɲ	ŋ	
	Voiceless Fricative		f	θ	s	ʃ		x	h
	Voiced Fricative		v	ð	z			ɣ	
	Trill			r					
	Lateral			l					
	Glide	w					j		
	Affricate				tʃ	dʒ			

are presented and grouped by vowel backness and vowel height while diphthongs are grouped by vowel backness. The consonants are categorized by manner of articulation and place of articulation as shown in Table II. Many consonants such as /b/, /p/, /m/, /f/, /s/ and so on are pronounced almost the same way as in English.

B. Malay Syllable Structures

When speaking, we can pronounce an isolated word in a sequence of smaller linguistic units called syllables. Syllabification is a useful way to segment speech and project rhythm. The internal structure of syllable include two components: onset and rhyme; rhyme of a syllable consists of a nucleus and an optional coda. Onset and coda are always consonants and nucleus is a vowel in most case. Onset is preceding nucleus while coda is following nucleus.

Most of the Malay syllables are in the form of Consonant-Vowel (CV) and Consonant-Vowel-Consonant (CVC) structure. Linguist usually consider Malay as a type III language [12], which has CV(C) structure. In this structure, coda is optional, and each syllable requires an onset and nucleus. An onset can have one or more consonants.

III. HMM-BASED TTS FOR MALAY

A. HMM-based Speech Synthesis Framework

The block diagram of the HMM-based speech synthesis framework is illustrated in Figure 1. The data-driven methodology is followed since the model is trained on a pre-recorded and annotated speech database. Thus, the framework consists of both a training part and a synthesis part.

In the training part, parametric representations of speech such as spectral and excitation parameters are extracted from a speech corpus. Then, these parametric representations are modeled by context-dependent continuous HMMs. The model parameters are estimated in maximum likelihood (ML) sense as:

$$\widehat{\lambda} = \arg\max_{\lambda} p(\mathbf{O}|W, \lambda)$$

where λ are the model parameters, $\mathbf{O}$ are parametric representations of training speech, and W are full-context label sequences converted from transcriptions of the speech corpus. The model parameters estimation process is usually similar to that in speech recognition system by applying expectation-maximization (EM) algorithm[16].

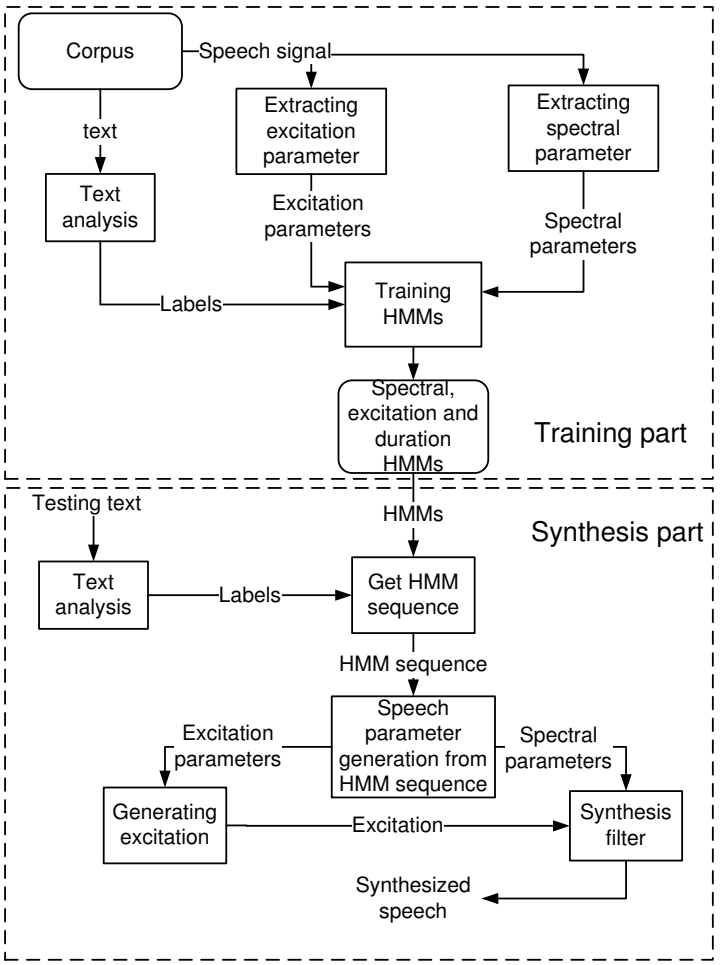

Fig. 1. HMM-based Text-to-Speech System Framework.

In the synthesis part, the input text is first converted to context-dependent labels using the text analysis module. Then, individual context-dependent HMM models are concatenated to represent the input utterance. From the model parameters $\widehat{\lambda}$ and context-dependent labels converted from input text w, parametric representations of an utterance $\mathbf{o}$ can be generated

by maximizing the output probabilities as follows:

$$\widehat{\mathbf{o}} = \arg\max_{\mathbf{o}} p(\mathbf{o}|w, \widehat{\lambda})$$

speech parameter generation algorithms proposed in [1] are applied to generate parametric representations.

Finally, speech waveforms are reconstructed from the generated parametric representations by applying vocoder techniques.

B. Contextual information

To capture the speech co-articulation effects and get better acoustic model, phonological, linguistic information and utterance structure are analyzed and then applied in the contextual factor construction when developing the system. The contextual factors at different linguistic level that have been extracted and taken into account are listed as following:

- Phonetic Level
  - The phoneme identity
  - The identities of the previous two and the next two phonemes (quin-phone)
  - The position of the current phoneme in the current syllable both forward and backward
  - The position of the current phoneme in the current word both forward and backward
- Syllable Level
  - Number of phonemes in previous, current and next syllables
  - The position of the current syllable in the current word both forward and backward
  - The position of the current syllable in the current utterance both forward and backward
- Word Level
  - Part-of-speech (POS) of previous, current and next words
  - The position of the current word in the current utterance both forward and backward
  - Number of phonemes in previous, current and next words
  - Number of syllable in previous, current and next words
- Utterance Level
  - Number of words in current utterance
  - Number of syllable in current utterance

IV. Experiments

A. Corpus Collection

A phonetically balanced text corpus is prepared by collecting the sentences from news websites. Each sentence consists of 5 to 10 words. From the text corpus, 1,350 sentences are used for speech recording. The speech corpora are carefully recorded under noise free condition. The waveform files are recorded using a headset microphone Sennheiser PC151, with mono channel analog-to-digital conversion at 16,000 Hz sample frequency. A female and a male native Malay speakers are asked to read the 1,350 sentences in broadcast news style, separately.

A text analysis module as shown in section III is implemented to generate phonetic labeling for the corpora. And a continuous HMM acoustic model trained for the Malay automatic speech recognition (ASR) system [11] is used to perform forced alignment for the corpora.

B. Experimental Setup

The speech corpora developed in previous section are divided into two parts: 1,300 utterances are used as training set and 50 utterances as testing set for female and male TTS system, respectively.

To analyze natural speech to get the speech representation parameters or reconstruct speech from parametric representations of speech, STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed Spectrogram)[13] is utilized as a vocoder under the HMM-based speech synthesis framework. In the analysis phase, we obtain smooth spectra by applying pitch-adaptive spectral analysis and time-frequency domain smoothing, in an effort to remove the effects of the original pitch. In the synthesis phase, we use pitch synchronous minimum-phase impulse response overlap-add to synthesize speech from spectrum and excitation parameters[13]. In this framework, we can easily manipulate speech parameters such as vocal track length, pitch or speech rate without significantly degrading the voice quality.

To train HMM acoustic model for TTS system, line spectrum pair (LSP) is selected to represent spectrum, since LSP has the closest relevance to the natural resonances or the "formants" of a speech sound[6]. 40 orders LSPs and 1 order log-scale energy are obtained from smoothed spectrum analyzed by STRAIGHT. Pitch is represented by log-scale fundamental frequency (F0) with their delta and delta-delta coefficient. F0s are automatically extracted on a short-time basis by applying the robust algorithm for pitch tracking[14]. When extracting F0, the speech signal is windowed by 25ms window with a 5ms shift for analysis. The feature vectors consisted of 40 orders LSPs, log-scale energy, log-scale F0s and their delta and delta-delta features. The inclusion of dynamics (delta and delta-delta features) ensures a smooth speech generation.

C. HMM model training

For acoustic model training, five-state, left-to-right without state-skipping HMM structure with single, diagonal Gaussian distribution is used to model spectral part. As far as F0 is concerned, F0 only exists in voiced segments and in unvoiced segments, no F0 is observed, multi-space probability distribution HMM (MSD-HMM)[15] is adopted to model F0 in this experiments.

The HMM model training procedure follows that introduced in [17]. In the HMM model training process, to capture the speech co-articulation effects in longer speech units (e.g. word, utterance), rich contextual factors are considered. Here, the HMM-based speech synthesis system results a total number of 37,546 rich context models (full-context HMM models).

TABLE III
LEAF NODE NUMBER OF DECISION TREE FOR LSP, F0 AND DURATION MODELS

	LSP	F0	Duration
Male	1139	2656	484
Female	1640	3319	562

As insufficient training data is available to model all context dependency, clustering is performed to group full-context HMM models so that the models are more robust for the unseen contexts in testing data. HMM states are clustered or tied via decision trees. Minimum description length (MDL)[18] criterion is used to control the model size when performing tree-based clustering. And MDL value is set to 1.0. The leaf nodes number of decision tree for LSP, F0 and duration models after clustering are presented in table III for reference.

D. Evaluation

To evaluate the developed HMM-based Malay TTS system, subjective evaluation is conducted. The subjective evaluation is based on the Mean Opinion Score (MOS), concerning the naturalness which describes how closely the synthesized speech like natural speech and the intelligibility which refers to the understandability of synthesized speech. The MOS score is on a scale of one to five where one stands for "bad" and five stands for "excellent". The evaluation is conducted as follows: we first present the original speech to the subjects as a reference, assuming the original speech is "excellent"; then the subjects listen to the synthesized speech. Subjects are asked to express their opinion for each sentence in a MOS scale. Each utterance can be heard more than once by the listeners.

In the evaluation test, 30 utterances pairs are used, which are not included in the training database; and each utterance is about 5 to 10 words long. A group of 4 listeners, who are native Malay speakers, participated in the subjective evaluation test and they were told to pay attention to the naturalness and the intelligibility when judging the speech quality.

TABLE IV
MOS OF SUBJECTIVE EVALUATION OF MALAY TTS SYSTEM

System	NATURALNESS	INTELLIGIBILITY
Male	3.2	4.3
Female	3.4	4.2

The subjective evaluation results are presented in Table IV, from which we can see the results are rather positive.

V. CONCLUSIONS

HMM-based speech synthesis provides a generalized statistical framework for speech synthesis. It is efficient for parametric speech modeling and speech parameter generation. In this work, HMM-based speech synthesis framework is utilized to develop the text-to-speech synthesis system for Malay language. Subjective evaluations on naturalness and the intelligibility aspects of the synthesized speech are performed to evaluate the Malay TTS system. The MOS results show that the developed HMM-based Malay TTS can synthesize speech with good quality.

VI. ACKNOWLEDGEMENTS

The authors would like to thank Ms Sharifah Mahani Aljunied from Institute for Infocomm Research for helpful discussion on Malay language linguistic knowledge, thank Dr. Xiao Xiong from Nanyang Technological University for providing the Malay dictionary and G2P program and all the listeners participated in the listening test .

REFERENCES

[1] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi and T.Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *in Proceedings of the IEEE International Conferance on Acoustics, Speech and Signal Processing (ICASSP),* vol. 3, 2000.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," *In Proceedings of Eurospeech ,* pp.2347-2350, 1999.

[3] K.F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *Readings in speech recognition,* Morgan Kaufmann, 1990.

[4] H. Zen, K. Tokuda and A.W. Black, "Statistical parametric speech synthesis," *Speech Communication,* vol.51, p.11, pp.1039-1064, Elsevier, 2009.

[5] K. Tokuda, H. Zen and A.W. Black, "An HMM-based speech synthesis system applied to English," *in Proceedings of the IEEE Speech Synthesis Workshop,* 2002.

[6] Y. Qian, F.-K. Soong, Y. Chen, and M. Chu, "An HMM-based Mandarin Chinese text-to-speech system," *In Proceedings of ISCSLP,* pp.223-232, 2006.

[7] S. Chomphan and T. Kobayashi, "Implementation and evaluation of an HMM-based Thai speech synthesis system," *In Proceedings of Interspeech,* pp.2849-2852, 2007.

[8] T. Drugman, A. Moinet and T. Dutoit, "On the use of machine learning in statistical parametric speech synthesis," *In Proceedings of Benelearn,* 2008.

[9] S. Krstulovic, A. Hunecke, M. Schroder, D. GmbH, and G.Saarbrucken, "An HMM-based speech synthesis system applied to German and its adaptation to a limited set of expressive football announcements," *In Proceedings of Interspeech,* 2007.

[10] O. Abdel-Hamid, D.Abdou and M.Rashwan, "Improving Arabic HMM-based speech synthesis quality," *In Proceedings of Interspeech,* pp.1332-1335, 2006.

[11] T.-P. Tan, H. Li, E.K. Tang, X. Xiao and E.S. Chng, "MASS: A Malay Language LVCSR Corpus Resource," *the 12th Oriental COCOSDA Workshop,* 2009.

[12] B.S. Teoh, "The sound system of malay revisited," *Kuala Lumpur: Dewan Bahasa dan Pustaka,* 1994.

[13] H. Kawahara, I. Masuda-Katsuse and A. de Cheveign, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0extraction: Possible role of a repetitive structure in sounds," *Speech Communication,* vol. 27, pp.187-207, 1999.

[14] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis,* vol. 495, Amsterdam, NL: Elsevier Science, 1995.

[15] K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS E SERIES D,* vol. 85, pp.455-464, INSTITUTE OF ELECTRONICS, INFORMATION, 2002.

[16] J.A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," *International Computer Science Institute,* vol. 4, 1998.

[17] K. Tokuda and H. Zen, "Fundamentals and recent advances in HMM-based speech synthesis," *Interspeech tutorial,* 2009.

[18] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *Acoustical Science and Technology,* vol.21, p.2, pp.79-86, 2000.