PAPER

Developing an HMM-Based Speech Synthesis System for Malay: A Comparison of Iterative and Isolated Unit Training

Mumtaz Begum MUSTAFA[†a)], ***Member***, **Zuraidah Mohd DON**[††], **Raja Noor AINON**[†], **Roziati ZAINUDDIN**[†††], ***and*** **Gerry KNOWLES**[††††], ***Nonmembers***

SUMMARY The development of an HMM-based speech synthesis system for a new language requires resources like speech database and segment-phonetic labels. As an under-resourced language, Malay lacks the necessary resources for the development of such a system, especially segment-phonetic labels. This research aims at developing an HMM-based speech synthesis system for Malay. We are proposing the use of two types of training HMMs, which are the benchmark iterative training incorporating the DAEM algorithm and isolated unit training applying segment-phonetic labels of Malay. The preferred method for preparing segment-phonetic labels is the automatic segmentation. The automatic segmentation of Malay speech database is performed using two approaches which are uniform segmentation that applies fixed phone duration, and a cross-lingual approach that adopts the acoustic model of English. We have measured the segmentation error of the two segmentation approaches to ascertain their relative effectiveness. A listening test was used to evaluate the intelligibility and naturalness of the synthetic speech produced from the iterative and isolated unit training. We also compare the performance of the HMM-based speech synthesis system with existing Malay speech synthesis systems.

key words: *iterative training, isolated unit training, cross lingual approach, uniform segmentation, segment-phonetic labels*

1. Introduction

Speech technologies are now available for many languages, and they include a number of useful systems and interactive tools, such as automatic speech recognition (ASR), speech synthesis and spoken dialogue systems. Their development is however uneven, and far out of reach for some languages. While there are approximately 6,000 different languages in existence [1], many of them do not have the continual development in speech technology enjoyed by more established languages such as English, Japanese, and few others, which have the advantage of sufficient resources to make continuing advances in speech synthesis. The term under-resourced language refers to languages that do not have adequate resources such as speech database, segment-phonetic labels that contain the time-alignment information of recorded speech, linguist experts and funding [2].

Many of these under-resourced languages, which are spoken in developing countries, including Bengali, Malay and Vietnamese, are amongst the 20 most widely spoken languages in the world [2]. The development of speech synthesis systems and other speech-related technologies for these languages will not only benefits people from these countries in the form of speech-related applications, but also enable comparisons to be made with other languages worldwide with respect to the influence of language structure on speech technology.

The lack of development of speech synthesis systems for under-resourced languages is more apparent in the case of the newer statistical parametric systems based on Hidden Markov Models (HMMs). The HMM-based speech synthesis systems have the ability to synthesize speech with a high degree of naturalness comparable to state-of-the-art unit selection systems [3]. The concept was first proposed by Yoshimura et al., [4] and was developed for languages such as Japanese, English, Thai, Romanian, Mandarin, Korean, Austria, Portuguese, Arabic, Hungarian and German among others [5]–[15].

HMM-based speech synthesis can be broadly divided into two parts: training and synthesis [3]. The training of HMMs is the process of creating the speech acoustic model based on speech parameters extracted from recorded speech. The synthesis part makes use of the speech acoustic model to generate the speech of an arbitrary text input [3], [4]. The naturalness of the synthesized speech therefore depends on the speech acoustic model built during training of the HMM.

There are two major approaches for training HMMs, namely iterative and isolated unit training [16]. The main difference of the two is that the former did not require any segment-phonetic labels [16] for training HMMs. Therefore, iterative training enables quicker development of speech acoustic model. This makes iterative training very attractive for building the speech acoustic model for a new and resource-constrained language. However, several researchers found that the speech acoustic model from iterative training produces synthetic speech which is less natural than isolated unit training [2], [16]–[21].

Iterative training is where the boundaries resulting from the previous iteration is used to initialize and re-estimate the speech acoustic models via the Baum-Welch algorithm [19]. A common algorithm for performing iterative

Manuscript received July 10, 2013.
Manuscript revised December 27, 2013.

[†]The authors are with the Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia.
[††]The author is with the Faculty of Languages and Linguistics and Asia-Europe Institute, University of Malaya, Malaysia.
[†††]The author is with the Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia.
[††††]The author is with the Lingenium Sdn Bhd, Kuala Lumpur, Malaysia.
a) E-mail: mumtaz@um.edu.my
DOI: 10.1587/transinf.E97.D.1273

training is Expectation Maximization Algorithm (EM) but the more recent algorithm for iterative training is the Deterministic Annealing EM (DAEM), which was introduced to overcome the problem of EM to minimize unreliable model parameters. Itaya et al. [20] concludes that the DAEM outperforms EM as it provides a simple iterative procedure to obtain appropriate maximum likelihood (ML) estimates.

Isolated unit training is performed using segment-phonetic labels [17]. However, for many under-resourced languages, access to segment-phonetic labels is virtually impossible. This is because the most tedious tasks in the resource accumulation process in the development of speech synthesis system are the segmentation and labeling [18]. In addition, manual segmentation has been criticized by several researchers because it tends to produce inconsistent results [16], [18].

In this paper, we explore several techniques for the development of an HMM-based speech synthesis system for an under-resourced language that lacks in recorded speech database and segment-phonetic labels. We are considering two forms of training HMMs which are iterative (benchmark) and isolated unit training. The purpose of this study is to show that we perform isolated unit training with little effort for an under-resourced language by making use of the resources developed for other more established languages. More specifically, we explore the possibilities of developing an initial model for an under-resourced language by adopting the speech acoustic model created for another language for automatic segmentation.

This paper is organized as follows: Section two gives an overview of Malay as under-resourced language. Section three describes the development of speech acoustic model for Malay, which includes the process of building the resources, training HMMs, and synthesizing speech. Section four explains the evaluation carried out on the synthetic speech generated by the newly developed HMM-based speech synthesis system in comparison to the existing Malay TTS systems. Section five discusses the findings, and section six concludes the paper.

2. Malay as an Under-Resourced Language

Based on [2], Malay can be considered as a partially under-resourced language when it comes to the development of HMM-based speech synthesis system including resources such as segment phonetic labels, experts (linguist) and the researchers needed for such developments.

One of the leading solutions for tackling resource issues for preparing segment phonetic label is the cross lingual approach, which provides a means of developing a speech synthesis system for a resource poor language, using the resources and model from a resource rich language. The cross-lingual approach has been successfully applied in the area of ASR system and segmentation of recorded speech of resource poor languages [2], [21], [22].

For cross-lingual adaptation, the context-dependent acoustic models developed for one language can be borrowed for use with another language. However, not much research has been carried out into cross-lingual context-dependent acoustic modeling to evaluate the effectiveness of such an approach. One problem that may arise from cross-lingual context-dependent acoustic model is context mismatching among different languages [2].

Because the cross lingual approach involves two different set of languages, it requires familiarity with orthography, phonology and morphology of both languages. An important aspect of the cross-lingual approach is to equate the sound from different languages, and in the absence of exact matches, select appropriate approximations [21], [22]. Identifying and exploiting a common inventory of sounds across languages enable the speech acoustic model of the source language to be used in the automatic segmentation of the target language.

It is therefore crucial to find a source language that have some form of similarity to the target language. However this does not mean that cross-lingual approach cannot use two languages with little similarity, where a phoneme mapping technique can be applied to map the phonemes of source and target language [23].

2.1 About Malay

Malay belongs to one of the western branches of the Austronesian language family, and it is widely spoken by more than 150 million people in Malay-speaking countries such as Malaysia, Indonesia, Brunei, Singapore and Southern Thailand. In Malaysia, Malay is spoken in many different dialects in different states. The focus of this research is on the standardized form of Malay. The term *Standard Malay* (SM) is generally taken to refer to the national norm or prestige dialect, which is also used as the official language in Malaysia [24], [25].

Although Malay in Malaysia and Malay in Indonesia (Bahasa Indonesia) have similar origins, the speakers of Malaysian Standard Malay in Peninsular Malaysia tends to speak at a more flowing pace, while words that end with the letter "a" often come out as a schwa (/ə/). Indonesian speakers speak in clipped staccato tones, their "r"s are more markedly trilled (rolled r), and nearly all words are pronounced exactly as they are spelled [26].

There are some common features between the Malay and English [27]. Like English, modern Malay uses the Roman alphabet for the written form, which was introduced during the British occupation of Malaysia. This greatly simplifies the task of predicting spoken forms from written ones.

Malay is popularly known as a phonetic language, which means that the pronunciation of Malay words can to a large extent be predicted from the spelling. English and Malay share several similarities in lexis, phonemes and categories of phoneme types. Table 1 compares English and Malay phonemes using the International Phonetic Alphabet (IPA).

Malay has 27 consonants, six vowels and three diph-

Table 1 Similarity of English and Malay Phonemic Inventory.

Phoneme type	IPA	Word Example in English	Example of Malay word
Vowels	[a]	as in the word '**a**re'	***a**yam* 'chicken'
	[e]	as in the word '**e**levated'	***e**jaan* 'spelling'
	[ə]	as in the word 'th**e**'	***e**mak* 'mother'
	[i]	as in the word 'c**ei**ling'	***i**lmu* 'knowledge'
	[o]	as in the word '**o**ld'	***o**mbak* 'wave'
	[u]	as in the word 's**u**per'	***u**ntuk* 'for'
Diphthongs	[ai]	as in the word 'p**ie**'	*bag**ai*** 'like'
	[au]	as in the word 'fall**ou**t'	*kal**au*** 'if'
	[oi]	as in the word 'b**oy**'	*amb**oi*** 'wow'
Nasal	[m]	as in the word '**m**adam'	***m**asam* 'sour'
	[n]	as in the word ' **n**ew'	***n**ama* 'name'
	[ɲ]	not available	*me**n**yala* 'aflame'
	[ŋ]	as in the word 'si**ng**'	*bu**ng**a* 'flower'
Fricatives	[s]	as in the word '**s**aw'	***s**atu* 'one'
	[h]	as in the word '**h**otel'	***h**abis* 'finish'
	[x]	as in the word 'lo**ch**'	***kh**idmat* 'service'
	[f]	as in the word '**f**ilm'	***f**ilem* 'film'
	[ʃ]	as in the word '**sh**e'	***sy**aitan* 'devil'
	[v]	as in the word 'violet'	***v**an* 'van'
	[z]	as in the word 'zebra'	*jena**z**ah* 'corps'
Affricates	[dʒ]	as in the word 'j**o**b'	***j**ari* 'finger'
	[tʃ]	as in the word 'c**h**eese'	***c**etek* 'shallow'
Plosives	[b]	as in the word 'b**u**gs'	***b**ola* 'ball'
	[p]	as in the word 's**p**y'	***p**eta* 'map'
	[d]	as in the word '**d**o'	***d**ari* 'from'
	[t]	as in the word '**t**ime'	***t**epat* 'precise'
	[g]	as in the word '**g**lass'	***g**elap* 'dark'
	[k]	as in the word ' s**k**y'	***k**alah* 'lost'
	[ʔ]	not available	*ena**ʔ*** 'nice'
Approximants	[w]	as in the word '**w**et'	***w**arna* 'colour'
	[j]	as in the word 'y**e**s'	***y**akin* 'confidence'
	[r]	as in the word '**r**ed'	***r**umah* 'house'
Lateral	[l]	as in the word '**l**ook'	***l**emah* 'weak'

thongs [24] all of which has its English counterparts except for the palatal nasal spelt "ny" and the glottal stop. Shadini and Rahim [28] found that there is phonemic similarity between Malay and English.

As in English, the main syllable structures in Malay are Consonant-Vowel (CV) and Consonant-Vowel-Consonant (CVC) [26]. These occur in almost every Malay primary word [29]. Other common syllable structures in Malay are Vowel (V) and Vowel-Consonant (VC).

Malay and English also share some differences and among the important differences is the English use of stress to emphasize a particular syllable or a particular word in a phrase to highlight the meaning. English is traditionally said to be a stress-timed language, whereas Malay is a syllable timed language [30]. In English the lengthening effect that accompanies the end of a phrase before a pause is regarded as a feature of the final word, in Malay it is a feature of the whole speech interval. Here, durations are markedly increased toward the end of speech units, but the rate of deceleration is not fixed or constant, so that duration at the end of a phrase is unpredictable.

Although the lack of resources causes problems for researchers working on Malay, a number of speech synthesis systems have been developed, including formant-based systems such as Sintesis Ucapan Melayu [31] and the Standard Malay Text to Speech System or SMaTTS [32]. Also developed is FASIH, a Malay diphone-based concatenative speech synthesis system based on the MBROLA synthesizer engine and NuSuara[†], a commercially available Malay speech synthesis system based on triphone unit selection.

[†]The NuSuara homepage is at http://www.nusuara.com

3. Building the Speech Acoustic Model of Malay

In this research, the speech acoustic model for Malay is built by training HMMs with two alternative techniques, which are iterative and isolated unit training.

3.1 Malay Speech Corpora

A Malay neutral speech database or MNSD [33] was constructed specifically for the development of the HMM-based speech synthesis system. The MNSD consists of 1,000 recorded utterances uttered by two Malay native speakers, one male and one female. The 1,000 Malay sentences were constructed specially for the recordings, and contain a representative range of Malay words, syllables and phones. They are also free of grammatical mistakes, and they were taken from various written sources such as local Malay newspapers (43%), educational text books (39%), and other general reading materials (18%). The selection of the sentences was made by a Malay linguist to ensure the phonetic richness of the database.

These 1,000 sentences include 588 short sentences (less than seven words) and 412 long sentences (seven words or more), with a range from three to 12 words per sentence, and an average length of about 5.5 words per sentence. The MNSD contains 5,534 word tokens (2,763 different word types), 12,666 syllables and 39,996 phones.

During the recording, the speakers were requested to speak in a neutral tone with little variation. The database contains 2.15 hours of recordings, made up of 1.04 hours for the male speaker and 1.11 hours for the female speaker. Recordings were made in studio conditions to minimize background noise. The microphone was placed at the standard 30 cm from the speaker's mouth. The sampling rate for all recordings was 44100 Hz (16 bit), and the files were saved in wav format.

3.2 Segment-Phonetic Label for Malay

The segment-phonetic labels for the HMM-based speech synthesis system [17] require a number of phonetic and linguistic contexts including the phone duration information, grapheme-to-phoneme (G2P) conversion and part of speech (POS) tagging for each phoneme. Table 2 shows the 39 contexts used to create the Malay labels.

Preparing the segment-phonetic labels manually is expensive in terms of both time and manpower. To simplify the task of preparing the segment-phonetic labels, a context-dependent label generating unit was developed to generate the phonetic and linguistic contexts of Malay.

Phone duration information of the segment-phonetic labels was derived from the automatic segmentation of the MNSD database using HTK automatic segmentation tools, since these have been shown in a number of developments to provide reliable segmentation [19]. HTK however requires

Table 2 The required contexts for Malay.

Level	Context number	Description
Phoneme	1, 2, 3, 4 and 5	Preceding of last, last, current, next and succeeding of next phoneme identity
	6 and 7	Location of the current phoneme in current syllable: forward and backward
	8, 9 and 10	Number of phoneme of previous, current and next syllable
Syllable	11 and 12	Position of the syllable in the current word: forward and backward
	13 and 14	Position of the syllable in the utterance: forward and backward
	15	Name of the vowel of the current syllable
	16, 17 and 18	The number of syllables in the previous, current and next word
	19, 20 and 21	Number of syllables in the previous, current and next phrase
	22	Number of syllables in this utterance
Word	23, 24 and 25	Gpos (guess part-of-speech) of the previous word, current word and next word
	26 and 27	Position of the current word in the current phrase: forward and backward
	28 and 29	Number of content words before and after the current word in the current phrase
	30	Number of words from the previous content word to the current word
	31	Number of words from the current word to the next content word
	32, 33 and 34	Number of words in the previous, current and next phrase
	35	Number of words in this utterance
Phrase	36 and 37	Position of the current phrase in utterance: forward and backward
	38	TOBI end tone of the current phrase
	39	Number of phrases in this utterance

a speech acoustic model for a particular language. Since Malay is one of a number of under-resourced languages for which a speech acoustic model is not yet available, we applied two approaches for the model initialization; uniform segmentation and supervised cross-lingual adaptation.

3.2.1 Uniform Segmentation Approach

For the uniform segmentation approach, the model initialization for the automatic segmentation of the Malay speech database was based on a fixed duration associated with each identifiable phoneme including pauses and silences. The model initialization for uniform segmentation is based on an EM technique, for which we applied 50 iterative training passes for each state. Although the iteration can be set at a much higher number, we set the limit at 50 iterations to save the computational cost of further iterative training.

3.2.2 Cross-Lingual Adaptation Approach

The cross-lingual adaptation proposed in this research transforms the acoustic model of English to represent the acoustic model of Malay. The source model for the adaptation consists of a speaker dependent English speech acoustic model obtained from the CMU ARCTIC[††] speech database, a U.S. English database (single-speaker), containing a total of 1132 phonetically balanced utterances. The speech acoustic model (context dependent HMMs and the state duration model) for English is adapted using 50 hand-labeled segmentations of Malay utterances using the Maximum Likelihood Linear Regression (MLLR) technique [2], [22], [34]–[37].

The number of sentences to be used for model adaptation was based on earlier work including work on speaker adaptation [38] and emotion adaptation [39]. The 50 utterances contain 2,743 phoneme tokens (35 phoneme types including silence and pause), and range in length from four to 11 words. The selected utterances reflect the complete range of phonemes and syllable structures in the MNSD. We excluded loan words of English origin, and included more words containing the palatal nasal spelt "ny" and the glottal stop. This is to compensate for the lack of information on palatal nasals and glottal stops in the English data.

[††]Available at http://hts.sp.nitech.ac.jp/?Download, retrieved on May 2008.

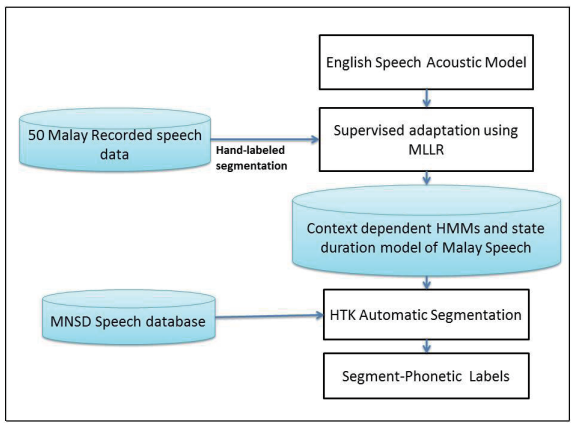

Fig. 1 The cross-lingual automatic phonetic segmentation of Malay speech using the English speech acoustic model.

The output of the automatic segmentation was then used to prepare segment-phonetic labels for training HMMs. Figure 1 shows the process involved in the cross-lingual automatic phonetic segmentation of Malay recorded speech initialized by the English speech acoustic model.

To map the phones of Malay and English, we used the data mapping approach proposed by Wu et al., [23] where the mapping information is used to attach the adaptation data of the target Malay speech to the source language model. In this method, the context dependent labels are mapped from the target language into the source language; in this case Malay segment-phonetic labels (which were prepared manually) are mapped into the English phonetic labels. The data mapping applied in this study deals with the issue of palatal /ɲ/ and glottal stop. For instance, the palatal /ɲ/ is mapped to the English speech acoustic model

as /n/+/y/, which is found in English words such as *onion*.

As the 36 Malay phonemes have approximate English counterparts (except for the palatal nasal /ɲ/ and the glottal stop), we have adapted the context related questions for decision tree of English (questions on contextual factors including phone identity and locational factors) with some modifications to the existing set of questions of English including the additional context of the palatal nasal /ɲ/ and glottal stop [ʔ], and excluding English vowels and diphthongs that are not found in Malay. We also modified the context related question for English acoustic model by excluding the leaf nodes for stress and accent which are not applicable to Malay.

The HEAdapt function in HTK toolkit [19] was used for the automatic segmentation of the MNSD recorded sentences. The output of the forced alignment then applied for preparing the segment-phonetic labels.

3.3 Accuracy of the Automatic Segmentation of MNSD

To evaluate the effectiveness of the automatic segmentation of MNSD speech data initialized by uniform segmentation and the cross-lingual approaches, we have measured the accuracy of the automatic segmentation equal to 20 ms tolerance rates with respect to the manual segmentation as proposed in [18].

The segmentation accuracy is measured at phoneme level for a sample of 50 utterances (not the 50 utterances manually segmented for the model adaptation) selected randomly. Any segmentation mismatch that is not within the tolerance limit of 20 ms i.e. exceeding 20 ms with respect to the hand-labelled segmentation is considered as segmentation error. This tolerance is considered in [40] as an acceptable limit to produce synthetic speech of good quality. The segmentation error is measured for each phoneme using the following formula:

$$Segmentation\ Error, (in\%), D_e = \frac{D_m - D_o}{D_m} * 100\% \tag{1}$$

where, D_m is the phoneme duration derived from the manual segmentation and D_o is the phoneme duration from automatic segmentation.

It was found that the segmentation error for the uniform segmentation approach was higher than for the cross-lingual adaptation approach. The average segmentation error for uniform segmentation was 6.33% as against 3.02% for the cross-lingual approach. Figure 2 compares the segmentation error of each phoneme type for both approaches. Among the eight types of phonemes, consonants are associated with a greater degree of error than vowels for both approaches. Among the six types of consonants, fricatives and affricates have the highest duration error. A possible reason for this, particularly for uniform segmentation, is the inaccuracy of the speech acoustic model in dealing with aperiodic segments.

Though the automatic segmentation initialized by both approaches contain some boundary errors, these errors can be reduced when performing the training HMMs.

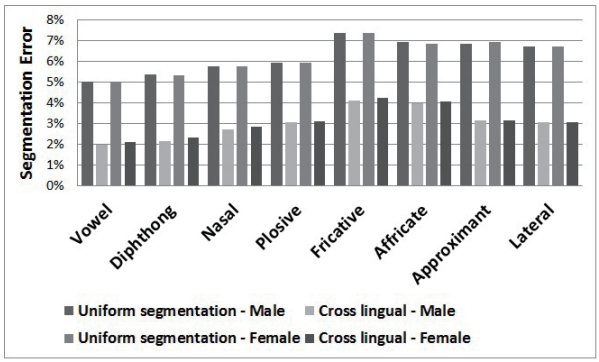

Fig. 2 Phoneme level duration segmentation error.

3.4 Training HMMs

To build the speech acoustic model for the synthesis of Malay speech, training HMMs was done using the MNSD speech database. To train the HMMs, we applied 950 utterances (excluding the 50 utterances segmented manually), in which all training data was sampled at 16 KHz and windowed using a 25-ms Blackman window with a 5-ms shift. The feature vectors for training consist of 25 mel-cepstral coefficients including the zeroth coefficient, the log F0, and their delta and delta delta coefficients. We applied the 5-state left-to-right HSMMs, and the spectrum of each state was modelled by a single diagonal Gaussian output distribution. The training made use of 38 phonemes including silences and pauses.

In this study, two forms of training HMMs were carried out, which are:

- Iterative training: using DAEM
- Isolated unit training: where the automatic segmentation using HTK toolkit was initialized using uniform segmentation and the cross-lingual approach (the training HMMs for the automatic segmentation using HTK was initiated by the segmental K-means algorithm, which is also known as Viterbi training [41]).

For the DAEM technique, the problem of maximizing the log-likelihood function is reformulated as the problem of minimizing a free energy functions as below:

$$F_\beta(\Lambda) = -\frac{1}{\beta}\log\int p(O,q|\Lambda)^\beta dq \tag{2}$$

We have applied the work in [20], in which the temperature parameter β was updated by applying the following formula:

$$\beta^{(i)} = \sqrt{i/I} \tag{3}$$

where $\beta^{(i)}$ is the value of β at i-th iteration, and I is the total number of the iterations. We used $I = 20$, and 10 iterations of the EM-steps were conducted at each temperature (which

Table 3 The leaf nodes of log F0, MGC and duration for Male voice.

State	Uniform segmentation approach			Cross-lingual approach		
	log F0	MGC	Duration	log F0	MGC	Duration
State 1	401	226	345	421	220	401
State 2	477	230	NA	446	239	NA
State 3	480	252	NA	460	236	NA
State 4	421	205	NA	466	227	NA
State 5	403	211	NA	524	206	NA
TOTAL	2182	1124	345	2317	1128	401

NA: The duration leaf node is only available for state 1

Table 4 The leaf nodes of log F0, MGC and duration for Female voice.

State	Uniform segmentation approach			Cross-lingual approach		
	log F0	MGC	Duration	log F0	MGC	Duration
State 1	656	236	473	716	284	528
State 2	553	223	NA	736	295	NA
State 3	529	231	NA	716	263	NA
State 4	651	247	NA	790	290	NA
State 5	526	210	NA	656	274	NA
TOTAL	2915	1147	473	3614	1406	528

NA: The duration leaf node is only available for state 1

means i ranges from 0^{th} iteration to the 10^{th} iteration per temperature).

For the isolated unit training with segmentation boundary made available from automatic segmentation, the speech database and labels that include initial phoneme segmentation were trained using 5 state left-to-right HSMMs with no skip. To begin with, monophone HSMMs are trained from the initial segmentation and construct untied full-context dependent HMMs. The full-context HSMM states were generated by introducing the phonetic, segmental, prosodic and linguistic context information.

We then performed Expectation Maximization (EM) re-estimation of the untied full-context dependent HMMs which is followed by state/stream clustering given the state alignment and the parameters of the untied model. From the 950 utterances, a total of approximately 1,800 unique models resulted from over 9,000 observed triphones. State tying was performed as it is a necessary condition for the training of state tied HSMMs. A single-pass re-estimation is performed during the estimation process which uses the tied models to get the state-level alignment of the training data.

Decision tree based state clustering was used for tying the full-context HMM states based on the minimum description length (MDL) principle. The tree is then applied to the HSMMs and the model parameters of the HSMMs are thus tied. The clustered HSMMs are re-estimated again. The clustering processes are repeated until convergence of likelihood improvements.

After that, we have performed several iterations of EM re-estimation of the clustered HSMMs which is then followed by the untied clustered HSMMs and perform one further EM re-estimation to get updated parameters of the untied full-context dependent HSMMs. This is followed with state/stream clustering given the state alignment and the parameters of the untied model by performing several iterations of EM re-estimation of the clustered HSMMs.

To illustrate the effect of applying different phoneme boundary (uniform segmentation and cross lingual) during the training HMMs, we compared the leaf nodes of log F0 (Fundamental Frequency), MGC (Mel-Generalized Cepstrum) and duration of both approaches for male and female speakers, as shown in Tables 3 and 4 respectively. It was found that the leaf nodes of the two approaches are different, those for the cross-lingual approach being higher than for the uniform segmentation approach. A larger number of leaf nodes for the cross-lingual approach shows that the HMM-based speech synthesis system has to handle greater variation in duration at phoneme level.

4. Listening Evaluation

We conducted two types of listening evaluation on the HMM-based synthetic speech. The first listening evaluation is to measure the intelligibility and naturalness of the synthetic speech generated by the HMM-based speech synthesis system built using different training HMMs. The second listening evaluation is to compare the naturalness of the HMM-based speech synthesis system developed in this research with existing Malay diphone-concatenative synthesis (TTS 1) and unit selection synthesis (TTS 2).

4.1 Evaluation of HMM-Based Synthetic Speech

In this listening evaluation, 50 semantically unpredictable Malay sentences (SUS) [42] were synthesized using the HMM-based speech synthesis system [17] for non-label training using DAEM and label training. These included five frequent syntactic structures:

- Subject - Verb - Direct Object (10 sentences)
- Adverbial - Transitive Verb - Direct Object (10 sentences)
- Subject - Verb - Adverbial (10 sentences)
- Question Word - Transitive Verb - Subject - Direct Object (10 sentences)
- Subject - Verb Complex - Direct Object (10 sentences)

The label generating unit was used to generate the required contexts for preparing the phonetic labels for the 50 SUSs. A total of 150 SUS was synthesized from all three speech acoustic models for Malay built from training HMMs.

50 native listeners of Malay took part in the listening evaluation, balanced for gender, age and profession. Listeners were asked to listen to a randomly provided sound folder containing 50 synthesized SUS in no particular order. We ensure that no folder contained the same SUS synthesized from two different speech acoustic models.

The intelligibility test was based on [8], and evaluators listened to the given set of utterances as many times as they wanted before typing in what they had heard. The intelligibility of the utterances was evaluated at word level, each correct word being given a score of 1 and incorrect words a score of 0. Typing errors and spelling mistakes were

ignored, so that spelling mistakes not affecting the meaning were not confused with intelligibility errors. For example, the mistyping of sembahyang 'pray', genggaman 'grip' or menunjukkan 'show' as "sembayang", "gengaman" or "menunjukan" was treated just as a spelling mistake and not an intelligibility error. For the Mean Opinion Score (MOS) [43] test, evaluators were asked to assess the naturalness of the synthetic voices on a scale from 1 'very unnatural' to 5 'very natural' as proposed in [44].

4.2 Comparative Evaluation of Different Malay Speech Synthesis Systems

For this listening evaluation, 20 sentences with an average length of 18.5 words were arbitrarily chosen from a leading Malay newspaper and synthesized by the existing TTS 1, TTS 2 and the HMM-based speech synthesis system. For the HMM-based synthesis, the context-dependent label generating unit was used to generate the phonetic labels for each sentence to be synthesized.

The comparative listening evaluation involved 50 Malay native evaluators not involved in the previous listening evaluation (which specifically for the testing of HMM-based speech synthesis system). The evaluators are demographically balanced in term of age, gender and profession. Each listener listened to 60 synthetic snippets (20 for each speech synthesis system in random order) and ranked them for intelligibility (a measure of how easily it can be understood), naturalness (degree to which the synthesized speech sounds like a human voice) and overall quality (freedom from distortion).

5. Results and Discussion

5.1 Evaluation of HMM-Based Synthetic Speech

Result of Intelligibility

The level of intelligibility achieved by the HMM-based system was found to be higher for the speech acoustic model built from isolated unit training HMMs (k-mean algorithm) than for the iterative training. However the intelligibility of uniform segmentation is poorer than the DAEM approach. Table 5 shows the intelligibility score for male and female synthetic speech.

The male voices were found to be more intelligible than the female voices, which could be due to a better quality of voice, consistent speech rate, F0 and higher energy (loudness) by the male speaker. We have performed the single factor ANOVA test to determine any significant differences of the mean intelligibility score.

The ANOVA test shows that there are significant differences of the mean intelligibility score for the three forms of training HMMs at $p < 0.05$ as shown in Table 6.

We have measured the intelligibility error of the male and female synthetic speeches generated using the different speech acoustic models. The total intelligibility errors for DAEM, uniform segmentation and the cross-lingual approaches were 4.60%, 4.81% and 3.41% respectively. Most of the WER is associated with longer words with many syllables. The WER is greater for words that vary in pronunciation or spelling according to local dialect, as opposed to standard words with standard spellings. Table 7 shows the WER according to word length for all approaches.

Table 5 The intelligibility score for male and female synthetic speech.

	DAEM	Uniform segmentation	Cross-lingual segmenation
Male	98.21%	98.04%	98.57%
Female	97.19%	97.15%	98.02%

Table 6 The results of the ANOVA tests for intelligibility score.

Comparison of mean	dF	F value	P value
DAEM and Uniform segmentation	1	22.22	0.000
DAEM and Cross lingual	1	852.44	0.000
Uniform segmentation and cross lingual	1	1402.39	0.000

Table 7 WER according to word length for all approaches.

Word	Acoustic model for Malay		
	DAEM	Uniform segmentation	Cross-lingual segmentation
1-syllable	0.31%	0.34%	0.21%
2-syllable	0.39%	0.45%	0.25%
3-syllable	0.57%	0.64%	0.32%
4-syllable	0.88%	0.91%	0.49%
>4-syllable	2.45%	2.47%	2.14%

Table 8 The naturalness score for male and female synthetic speech.

	DAEM	Uniform segmentation	Cross-lingual segmenation
Male	4.14	4.12	4.37
Female	4.04	4.03	4.14

Result of Naturalness Evaluation

Table 8 shows the naturalness score for male and female synthetic speech at the 95% confidence level. The naturalness of the speech generated by the cross-lingual approach was found to be better than both DAEM and uniform segmentation.

We have also performed the single factor ANOVA test for naturalness score. There are significant differences of the mean naturalness score for the three forms of training HMMs at $p < 0.05$. Table 9 shows the results of the ANOVA tests.

5.2 Comparative Evaluation of Different Malay Speech Synthesis Systems

The comparative evaluation of several Malay synthesis systems shows that evaluators ranked the HMM-based speech

Table 9 The results of the ANOVA tests for naturalness score.

Comparison of mean	dF	F value	P value
DAEM and Uniform segmentation	1	71.26	0.000
DAEM and Cross lingual	1	3686.43	0.000
Uniform segmentation and cross lingual	1	4498.49	0.000

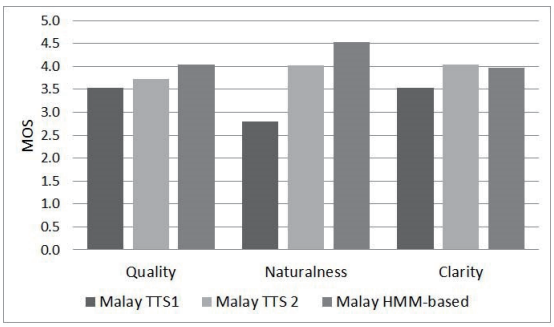

Fig. 3 Comparison of the HMM-based system with two other Malay TTS systems.

synthesis system highest for quality and naturalness, and TTS 1 the lowest. The commercially available TTS 2 obtained a slightly higher score for clarity than the HMM-based speech synthesis system. This is because TTS 2 generates synthetic speech with consistent pauses between words. HMM-based speech synthesis system on the other hand produces a continuous string of speech. This makes the HMM-based speech synthesis system's speech faster than the TTS 2, affecting the clarity but with greater naturalness than TTS 2, which the evaluators perceived as machine-generated speech in view of the long and consistent pauses between words. All measures of MOS are at 95% confidence level.

TTS 1 ranks lowest on all counts, and came far below TTS 2 and the HMM-based system for naturalness and clarity. Some evaluators commented that TTS 1 sounded very machine like, while TTS 2 lacked the smooth rhythmical flow that makes a voice sound natural. Most evaluators agreed that the HMM-based system came closest in naturalness to a human voice as reflected in its high score for naturalness in the evaluation. Figure 3 compares the clarity, naturalness and quality of the three TTS systems.

Native Malay listeners give a very high ranking for naturalness and intelligibility to the HMM-based system, which confirms that this system is able to generate high quality synthetic speech. The results of the listening test also show that the speech acoustic model generates better synthetic speech using English acoustic model than uniform segmentation. The comparison for naturalness of the HMM-based system developed in the course of this research with existing commercial TTS systems confirms the superiority of HMM-based synthesis for Malay.

6. Conclusions

Parametric speech synthesis systems, including those based on HMMs, can generate synthesized speech of acceptable quality. This form of synthesis is not yet available for many languages, as it still requires language dependent resources such as recorded speech and segment-phonetic labels. Under-resourced languages lack some or all language-dependent resources, making it difficult to develop an HMM-based speech synthesis system.

However, techniques such as iterative training and cross-lingual adaptation enable a development of HMM-based speech synthesis systems with minimal resources and effort.

Evaluators ranked the naturalness and intelligibility of synthetic voices produced by the isolated unit training, especially cross-lingual approach higher than the iterative training. On top of that, the comparison of duration shows that the cross-lingual approach generates durations closer to those in the recorded utterances compared to uniform segmentation approach. This is supported by the results of the ANOVA tests.

The superior performance of the cross-lingual approach is attributable to (1) the use of source data adapted from English CMU speech data, which had already been segmented with a high degree of accuracy, and (2) similarities between Malay and English. From these findings, it can be concluded that for HMM-based speech synthesis systems, the speech acoustic model build from isolated unit training with accurate segmentation synthesized a more natural speech than iterative training based on DAEM.

At the start of this research, there was no HMM-based speech synthesis system available for Malay in view of the scarcity of resources. The outcome of this research is that we have successfully developed a state-of-the-art speech synthesis system for Malay with an acceptable degree of intelligibility and naturalness based on HMMs by adopting the cross-lingual approach and making use of resources from a more established language, in this case English.

Acknowledgments

This work was supported by a research grant (Grant No.: RG019/09ICT) from University of Malaya, Malaysia.

The authors would like to thank the reviewers for their continuous effort in reviewing the manuscript and providing useful comments in order to improve the quality of this manuscript.

References

[1] T. Schultz and K. Kirchhoff, Multilingual speech processing, New York Academic Elsevier, p.536, 2006.

[2] V.B. Le and L. Besacier, "Automatic Speech recognition for under-resourced languages: Application to Vietnamese language," IEEE Trans. Audio Speech Language Process., vol.17, pp.1471–1481, 2009.

[3] H. Zen, K. Tokuda, and A.W. Black, "Statistical parametric speech synthesis," Speech Commun., vol.51, no.11, pp.1039–1064, 2009.
[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-Based speech synthesis," Proc. EUROSPEECH-1999, pp.2374–2350, 1999.
[5] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, Y. Guan, R. Hu, K. Oura, Y. Wu, K. Tokuda, R. Karhila, and M. Kurimo, "Thousand of voices for HMM-based speech synthesis-analysis and application of TTS systems built on various ASR corpora," IEEE Trans. Audio, Speech Language Process., vol.18, no.5, pp.984–1004, 2010.
[6] K. Tokuda, H. Zen, and A.W. Black, "An HMM-based speech synthesis system applied to English," Proc. IEEE Workshop on Speech synthesis, pp.227–230, 2002.
[7] S. Chomphan and T. Kobayashi, "Tone correctness improvement in speaker dependent HMM-Based Thai speech synthesis," Speech Commun., vol.50, pp.392–404, 2008.
[8] B. Stan, J. Yamagishi, S. King, and M. Aylett, "The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-Based speech synthesis system using a high sampling rate," Speech Commun., vol.53, pp.442–450, 2011.
[9] Y. Li, S. Pan, and J. Tao, "HMM-Based speech synthesis with a flexible Mandarin stress adaptation model," Proc. 10th ICSP2010 Proceedings, Beijing, pp.625–628, 2010.
[10] S.J. Kim, J.J. Kim, and M.S. Hahn, "Implementation and evaluation of an HMM-Based Korean speech synthesis system," IEICE Trans. Inf. & Syst., vol.E89-D, no.3, pp.1116–1119, March 2006.
[11] M. Pucher, D. Schabus, J. Yamagishi, F. Neubarth, and V. Strom, "Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis," Speech Commun., vol.52, pp.164–179, 2010.
[12] R. Maia, H. Zen, K. Tokuda, T. Kitamura, and F.G. Resende, Jr., "Towards the development of a Brazilian Portuguese text-to-speech system based on HMM," Proc. EUROSPEECH 2003, pp.2465–2468, 2003.
[13] O. Abdel-Hamid, S. Abdou, and M. Rashwan, "Improving Arabic HMM based speech synthesis quality," Proc. INTERSPEECH 2006, pp.1332–1335, 2006.
[14] B. Toth and G. Nemeth, "Hidden-Markov-model based speech synthesis in Hungarian," J. Info-Communication, vol.7. pp.30–34, 2008.
[15] S. Krstulovic, A. Hunecke, and M. Schroeder, "An HMM-based speech synthesis system applied to German and its adaptation to a limited set of expressive football announcements," Proc. INTERSPEECH-2007, pp.1897–1900, 2007.
[16] I. Mporas, A. Lazaridis, T. Ganchev, and N. Fakotakis, "Using hybrid HMM-based speech segmentation to improve synthetic speech quality" 13th Pan-Hellenic Conference on Informatics, pp.118–122, 2009.
[17] K. Tokuda, H. Zen, J. Yamagashi, T. Masuko, S. Sako, A.W. Black, and T. Nose, "HMM-based speech synthesis system (HTS) version 2.1," 2008, http://hts.sp.nitech.ac.jp/ (accessed and downloaded December, 2008).
[18] S. Jarifi, D. Pastor, and O. Rosec, "A fusion approach for automatic speech segmentation of large corpora with application to speech synthesis," Speech Commun., vol.50, no.1, pp.67–80, 2008.
[19] S. Young, G. Evermann, M. Gales, H. Thomas, D. Kershaw, and X. Liu, "The HTK Book (HTK Version 3.4)" Cambridge University Engineering Department, pp.1–349, 2006.
[20] Y. Itaya, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura, "Deterministic annealing EM algorithm in parameter estimation for acoustic model," Proc. INTERSPEECH-2004, pp.433–436, 2004.
[21] D.R.V. Nierkerk and E. Barnard, "Phonetic alignment for speech synthesis in under-resourced languages," Proc. INTERSPEECH-2009, pp.880–883, 2009.
[22] K.U. Ogbureke and J.C. Berndsen, "Framework for cross-language automatic phonetic segmentation," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2010), pp.5266–5269, Dallas, Texas, USA, 2010.
[23] Y.J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," Proc. INTERSPEECH-2009, pp.528–531, 2009.
[24] G.O. Knowles and M.D. Zuraidah, Word Class in Malay First ed. Kuala Lumpur, Dewan Bahasa and Pustaka, Malaysia, 2006.
[25] Y. El-Iman and M.D. Zuraidah, "Rules and algorithms for phonetic transcription of standard Malay," IEICE Trans. Inf. & Syst., vol.E88-D, no.10, pp.2354–2372, Oct. 2005.
[26] M. Yunus, The Malay Sound System, Kuala Lumpur, Fajar Bakti Sdn. Bhd. Malaysia, 1980.
[27] N. Seman, "Acoustic pronunciation variations modeling for standard Malay speech recognition," Computer and Information Science, vol.1, pp.112–120, 2008.
[28] A.H. Shahidi and R. Aman, "An acoustical study of English plosives in word initial position produced by Malays," The Southeast Asian Journal of English Language Studies, vol.17, no.2, pp.23–33, 2011.
[29] T.H. Nong, J. Yunus, and S. Hussain, "Speaker-independent Malay syllable recognition using singular and modular neural networks," Jurnal Teknologi, pp.65–76, 2001.
[30] M.D. Zuraidah, G. Knowles, and Y. Janet, "How words can be misleading: A study of syllable timing and 'Stress' In Malay," Linguistics Journal, vol.3, no.2, pp.66–81, 2008.
[31] H. Aini, A.S. Salina, and K.T. Soon, "Theory, methodology and implementation of the Malay text-to-speech system," Malaysian J. Computer Science (MJCS), vol.12, no.1, pp.28–37, 1999.
[32] O.K. Othman, Z.H. Ahmad, and T.S Gunawan, "SMaTTS: Standard malay text to speech system," Int. J. Comput. Science, vol.4, no.2, pp.285–293, 2007.
[33] B.M. Mumtaz, R.N. Ainon, R. Zainuddin, M.D. Zuraidah, and G. Knowles, "A cross-lingual approach to the development of an HMM-Based speech synthesis system for Malay," Proc. INTERSPEECH-2011, Florence, Italy, pp.3197–3200, 2011.
[34] B. Wheatley, K. Kondo, W. Anderson, and Y. Muthusamy, "An evaluation of cross language adaptation for rapid HMM development in a new language," Proc. ICASSP 1994, pp.1237–1240, 1994.
[35] C. Nieuwoudt and E. Botha, "Cross lingual use of acoustic information for automatic speech recognition," Speech Commun., vol.38, pp.101–113, 2002.
[36] M. Adda-Decker, L. Lamel, and N.D. Snoeren, "Initializing acoustic phone models of under-resourced languages: A case-study of Luxembourgish," 2nd Workshop on Spoken Languages Technologies for Under-resourced languages, pp.74–80, Penang, Malaysia, 2010.
[37] B.M. Mumtaz and R.N. Ainon, "Emotional speech acoustic model for Malay: iterative versus isolated unit training," J. Acoustical Society of America, vol.134, no.4, pp.3057–3066, 2013.
[38] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithm for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," IEEE Trans. Audio Speech Language Process., vol.17, no.1, 2008.
[39] M. Tachibana, J. Yamagashi, T. Masuko, and T. Kobayashi, "A style adaptation technique for speech synthesis using HSMM and suprasegmental features," IEICE Trans. Inf. & Sys., vol.E89-D, no.3, pp.1092–1099, March 2006.
[40] J. Matousek, D. Tihelka, and J. Psutka, "Automatic segmentation for Czech concatenative speech synthesis using statistical approach with boundary-specific correction," 8th European Conf. on Speech Communication and Technology, EUROSPEECH-2003, pp.301–304, 2003.
[41] B.H. Juang and L.R. Rabiner, "The segmental K-means algorithm for estimating parameters of hidden Markov models," IEEE Trans. Acoust. Speech Signal Processin vol.38, no.9, pp.1639–1641, 1990.
[42] C. Benoit and M. Grice, "The SUS test: A method for the assessment of text-to-speech intelligibility using semantically unpredictable sentences," Speech Commun., vol.18, pp.381–392, 1996.

[43] CCITT, Absolute category rating (ACR) method for subjective testing of digital processors, Red Book, 1984.
[44] S. King, K. Tokuda, H. Zen, and J. Yamagishi, "Unsupervised adaptation for HMM-based speech synthesis," INTERSPEECH-2008, pp.1869–1872, 2008.

Mumtaz Begum Mustafa received the BSc. and MSc. in Software Engineering from University Putra Malaysia (UPM) and University Malaya (UM) in 2002 and 2006 respectively, and the Ph.D. in Computer Science from University Malaya (UM) in 2012. She is currently a Lecturer at the Department of Software Engineering, University Malaya. Her research interests include Speech Synthesis and its applications. She has published her work in many of prestigious international speech conferences. She is a member of The Institute of Electronics, Information and Communication Engineers (IEICE) organization and the member of International Speech Communication Association (ISCA).

Zuraidah Mohd Don is a Professor at the Department of English Language, Faculty of Languages and Linguistics, University of Malaya. Her research interests include Prosody and Corpus Linguistics. Her recent articles have been published in Journal of Pragmatics, International Journal of the Sociology of Language, Linguistics Journal, Multilingual and International Journal of Corpus Linguistics. She is the member of Asia-Europe Institute, University of Malaya, Malaysia.

Raja Noor Ainon is an Associate Professor at the Department of Software Engineering, Faculty of Computer Science, and University of Malaya. Her research areas include Malay Text-to-Speech Synthesis, Multilingual Speech Recognition, Genetic Algorithms and Soft Computing. She is the author of more than 30 scholarly articles in Automatic Timetabling, Text compression, Expert Systems, Computational Linguistics, Genetic Algorithms, Emotional Text-to Speech Synthesis and Speech Recognition.

Roziati Zainuddin is a Professor attached to Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya. Her areas of interest are in intelligent multimedia, image and speech processing, computational fluid dynamics, biomedical informatics, computer vision visualization and E-Learning. Her research work has been published in several international journal and conference publications.

Gerry Knowles spent 2002-2003 as Visiting Professor in the Faculty of Languages and Linguistics, University of Malaya on sabbatical from Lancaster University in the UK. His research interests include corpus linguistics, phonetics and history of English. His most recent include two co-authored books entitled Malay Word Class: A corpus-based Approach and Malay Adverbs: Problems and Solutions. He is currently involved in several projects including the project on speech and the MALEX project.