High Quality Malay Speech Synthesis with Phonetic Balance Database Design for Embedded System

Lau Chee Yong, Tan TianSwee
Medical Implant Technology Group, Materials and Manufacturing Research Alliance,
Faculty of Biosciences and Medical Engineering,
UniversitiTeknologi Malaysia,
Skudai 81310, Johor, Malaysia
laucheeyong@hotmail.com; tantswee@biomedical.utm.my

Abstract:-Speech synthesis plays a pivotal role in consumer devices of our daily life. In this paper, we aim to propose an efficient Malay language speech synthesis system that conforms to the language's features. Firstly, a fully phonetic balance database; this database is then utilized to synthesize the speech using statistical parametric method. The proposed combination of the constructed database and the speech synthesis algorithm exhibits superiority in terms of footprint reduction attributable to the optimized constructed database while achieving relatively high quality in synthesized speech. The quality of synthesized speech is assessed by 60 evaluators; Quality Mean Score of 4.14/5 is obtained. The total file size of speech synthesizer is less than 2 Megabytes; the relatively smaller database size indicates high efficiency in footprint that in turn proves its feasibility and performance in embedded system of consumer devices. We conclude that we have constructed a Malay language speech synthesis that is favorable to be applied in embedded system in consumer devices on the ground of small footprint and considerably high synthesized speech quality based on Malay language features.

Key-Words: -Speech synthesis, hidden Markov model, phonetic balanced, footprint.

1 Introduction

Speech synthesis technique has come to prominence in many applications including speech rehabilitation such as talking head designed to assistpeople in language pronunciation learning, words' semantics understanding, and language sense fostering and self-awareness of language learning.The language training raises up certain ability to understand and speak, to listen and say, namely deaf but not mute, to establish a foundation in knowledge learning and social development [1].Besides, speech synthesis provides an important aid for blind or partially-sighted people, screen reading software is incorporated[2, 3].Rehabilitation purpose is by no means the only application of speech synthesis; speech synthesis is applicable in various context such as in our mobile phones, car navigation devices, human-machine interactive tools and so on [4, 5]. To embed speech synthesis into devices, low consumption of memory is preferred to ensure superior characteristics in terms of processing smoothness and response time at the expense of synthesized speech's quality. The trade-offs between the consumed memory's size and output speech's quality represents a considerable challenge to embedded speech synthesis technique [6].

In this study, a Malay language speech synthesizer based on statistical parametric method has been designed [7] using hidden Markov model (HMM). Its ability to extend to other languages is become a preference in speech technology study. However, only unit selection method has been used in designing a Malay language speech synthesizer currently [8]. So in this study we hope to extend the speech technology in Malay language by applying a newer technique which is HMM-based statistical parametric speech synthesis.

2 Problem Formulations

The selection of database for speech synthesizer is very crucial to construct a high quality speech. So in our study, we carefully design the corpus and not select the words randomly to reach a phonetic balance state in our database. This approach is to make sure every phoneme of Malay language were trained into the HMM and ensure the output speech quality.

Also, current Malay speech synthesizer available is big in total system size. It is because the database must be included in the system and the size of database is often large. This is not practical to be embedded into daily used devices. So in this study, statistical parametric speech synthesis method was adopted in Malay language to create a relatively smaller system size of speech synthesizer and it is more suitable to be embedded into daily used system.

In this study, a Malay language database was created by taking care of all phonetic samples.The script of database was designed suitable for daily application usage. The synthesizer was design to be smaller in total file size which is suitable to become an embedded system.

3 Problem Solutions

3.1 Database Construction

To construct a complete database for training, the script for the recording must be phonetically balance to provide all the phoneme type. In Malay language, there are 24 pure phonemes and 6 borrowed phoneme which can be categorized into 8 different groups. The table below lists down the phonemes according to its group.

Table1
List of Malay Phoneme According to Group

Category	Malay Phones
Vowels	/a/, /e/, /eh/, /i/, /o/, /u/
Plosives	/b/, /d/, /g/, /p/, /t/, /k/
Affricates	/j/, /c/
Fricatives	/s/, /h/, /f/, /z/, /sy/, /kh/, /gh/, /v/
Nasal	/m/, /n/, /ng/, /ny/
Trill	/r/
Lateral	/l/
Semi-vowel	/w/, /y/

Six borrowed consonantal phonemes are /f, z, sy, kh, gh, v/ and 5 diphthongs in Malay language are /ai/, /au/, /oi/, /ua/, /ia/. The constructed database consists of all these phonemes including silence, /pau/. The phoneme /gh/ has been folded into /g/ due to the similarity in pronunciation [9].

The script of the database was constituted only Malay words. The words were gathered from online Malay news such as BeritaHarian, Bernama, Cybersing, Malaysia Kini, Utusan and primary school Malay Language textbooks. The reason of finding words from newspaper and Malay Language primary school textbooks is to gather the basic words that are frequently used in our daily life. The gathered words from this source are unlikely to content special words that only can be found in certain fields. Therefore, it is suitable to be used in daily application. The duration of Malay words collection was 3 months and total of 10 million words have been collected. The texts from online news were in html format and a text processing tool was designed to extract the words automatically. The words were extracted by eliminating all the punctuation and detect words boundary by using space. Total of 115738 of different Malay words were collected.

The sentence design using all of the collected Malay words is less practical. Hence, from the collected words, a threshold value of 70% of highest frequency was chosen as a benchmark for the words selection. Total of 1451 words found in that range of frequency. The reason for choosing 70% is because the number of words collected is practical for training and covers a sufficient percentage of words (Tan and Salleh, 2008).The number of Malay words found according different percentage value is listed in the table below.

Table 2
Word Coverage According To Frequency of Occurrence

Category	Total words
60% frequency words	747
65% frequency words	1025
70% frequency words	1451
80% frequency words	2592

In the word collection, top 10 of highest frequency words wereconcluded. Thus in the sentence design, these 10 words were frequently added into database to ensure a better training result. The top 10 highest frequency words and their frequency of occurrence are summarized in the table below.

In total, 381 sentences were designed based on the words from online Malay news. Another 607 sentences were designed according to the words found in primary school Malay textbooks. The non-repeat words were extracted from textbooks and used to design the sentence. As a result, a fully phonetic balanced database script was designed which includes all the phoneme types and each type of phoneme occurrence is more than 5 times.

Table 3
Top 10 of Highest Frequency of Word Occurrence

Word	Frequency of Occurrence
Yang	282200
Dan	253097
Untuk	91374
Tidak	84443
Pada	60179
Akan	58222
Saya	55500
Kepada	55497
Mereka	55175
Ke	42310

3.2 Speech Training and Synthesis

Figure 1 shows the process of training and synthesis of speech. We adopted statistical parametric speech synthesis method introduced by [7]. Excitation and spectral parameters will be extracted [10]from database and trained into HMM. To estimate the model parameter, maximum likelihood criterion (ML) is usually used according to the equation below.

$$\hat{\lambda} = \arg\max_{\lambda} \{ p(O \mid \omega, \lambda) \} \qquad (1)$$

Where λ is a set of model parameters, *O* is a set of training data, and ω is a set of word sequences corresponding to *O*. We then generate speech parameters, *o*, for a given word sequence to be synthesized, *w*, from the set of estimated models, $\hat{\lambda}$, to maximize their output probabilities as

$$\hat{o} = \arg\max_{o} \{ p(o \mid w, \hat{\lambda}) \} \qquad (2)$$

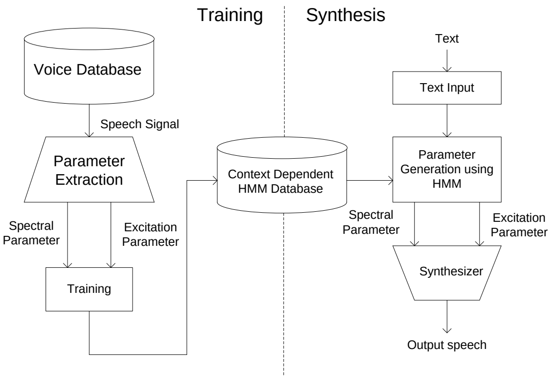

Fig. 1. Process of training and synthesis of speech. (Adapted from [7])

After that, a speech waveform is rebuilt from the parameter.Three important elements for HTS speech synthesis are mel-cepstrum, generalized log f0, and band limited aperiodicity measure. The training process manipulates these elements and result in updated model with parameters.

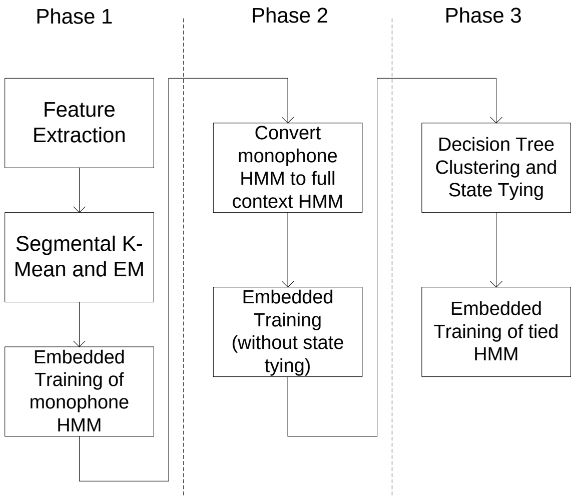
Phase 1
Phase 2
Phase 3
Feature Extraction
Segmental K-Mean and EM
Embedded Training of monophone HMM
Convert monophone HMM to full context HMM
Embedded Training (without state tying)
Decision Tree Clustering and State Tying
Embedded Training of tied HMM

Fig. 2. Block Diagram of Training Process

The overall training process can be described as three phases as illustrated in the figure above. At the first phase, the monophone HMMs are trained with the initial segments of database script and speechusing segmental k-means to classify the group of phonemes and Expectation-Maximization (EM) algorithm [11] was used to performembedded training of the monophone. At phase 2, monophone HMMs were converted into context dependent HMMs. Re-estimation was done using embedded training until the parameters were converged. At phase 3, the decision tree clustering [12]is applied for the spectral stream, log f0, band limited aperiodic measures and duration probability distribution functions (pdf).The tied models were further trained until the parameters converged.

At the synthesis stage, first the target text to be synthesized was converted into context-labeled sequence. Then, the corresponding context-dependent HMMs were concatenated to construct the target sentence. The state durations of the HMM were determined so as to maximize the probability of state durations [13]. And the mel-cepstral coefficients and log F0 values are determined using speech parameter generation algorithm like in Case 1 of [14]. By the inclusion of dynamic coefficients, the output speech is constrained to be realistic. Finally, the target speech waveform is synthesized directly based on the generated mel-cepstral coefficients and F0 values using STRAIGHT mel-cepstralvocoder with mixed excitation [15, 16]. The process can be illustrated as below.

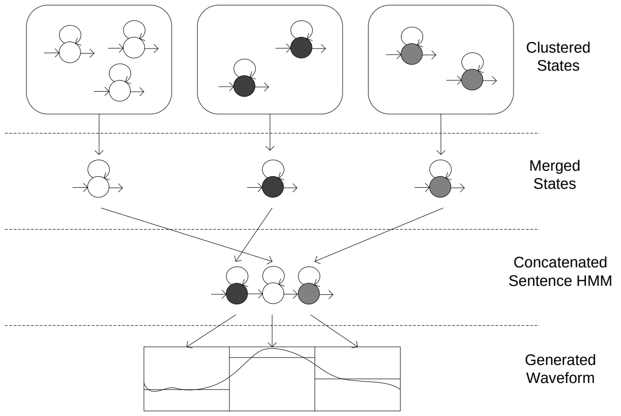

Fig. 3. Block Diagram of Speech Waveform Generation Process. (Adapted from [7])

4 Result and Discussion

Output speech was evaluated by 60 people. They were asked to listen to the synthetic voice and the original recorded voice with and without headphone. Rating was done according to their opinion about the similarity of synthetic voice with original voice, the naturalness and intelligibility of the synthetic voice. The result was analyzed using Statistical Product and Service Solution (SPSS) software. The result is summarized in the table and figure 4 as below.

In the graph, x axis represents the rating, 5 is the best while 1 is the worst. Y axis represents the number of listeners give the rating. From the result, the mean score for similarity is 4.43 which means majority of the listeners feels that the synthetic voice is similar to the original voice. This is due to the melcepstral order used was 39. So it captures a lot of sound features of the original voice hence the synthetic voice sounds similar to original voice. Listeners were asked to rate the naturalness of the synthetic voice without intelligibility concern. The mean score is 4.17denotes that they found the synthetic voice is natural. However, minority of them observed that the synthetic voice sounds rigid. This is because the time frame of the synthetic voice phoneme is almost the same so the style of the output voice sounds robotic. About the intelligibility, they were asked to rate if they understand the spoken speech. The mean score is 3.82shows that almost 70% of evaluators understand the synthetic speech.

According Table 5, the total footprint is less than 2Mbytes with no compression. By comparing the work done by [8], the file size of the Malay speech synthesizer is 37.6 Mbytes. Total reduction of 95.1% of file size was achieved in this study. So it is suitable to become an embedded system in devices [5, 7].

Table 4
Evaluation Test Result

	Similarity		**Naturalness**		**Intelligibility**	
Rating	Freq	%	Freq	%	Freq	%
1	0	0	0	0	0	0
2	0	0	2	3.3	0	0
3	5	8.3	5	8.3	21	35.0
4	24	40.0	34	56.7	29	48.3
5	31	51.7	19	31.7	10	16.7
Total	60	100	60	100	60	100
Mean	**4.43**		**4.17**		**3.82**	

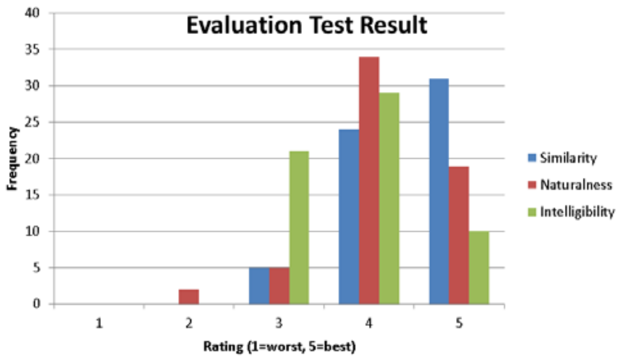

Fig. 4. Evaluation Test Result

Table 5
Footprint of Constructed System in kByte

Module		**Size (kByte)**
Decision Tree	Spectrum	114.0
	F0	298.0
	Duration	95.2
Probability Density Function	Spectrum	1126.0
	F0	151.0
	Duration	36.0
Converter		3.0
Synthesizer		36.0
Total		**1859.2**

The Malay language characteristics possess as advantages in this study. The pronunciation of the diphones of Malay language is straight forward and almost equally and constantly. Moreover, it is not a tonal language. Multiple training of same diphones with different tones is excluded. Besides, Malay language is written in Roman character which means the grapheme to phoneme conversion is also excluded in speech synthesis process[17].

From the result, the overall quality of the output speech is rated at 4.14 out of 5. This indicates that the database design is applicable for speech synthesis and able to obtain good quality synthetic speech with enough training data. By properly construct the database, the training sentence, recording time and training time can be much reduced.

In short, fully phonetic balance database can ensure the output speech quality while small footprint enables it to be implemented into embedded system.

5 Conclusion

This paper presented a Malay speech synthesis with proper design of database.A fully phonetic balanced database was constructed with more than 5 times of occurrence of each phoneme. The scripts of the database were composed for daily application usage. Statistical parametric speech synthesis method was utilized to create the synthesizer. As the result, the synthesized speeches were rated by 60 people and obtained average mean score of 4.43, 4.17 and 3.82 for similarity, naturalness and intelligibility respectively in total score of 5. To be embedded into devices, this speech synthesis system also possesses low footprint which is less than 2MByte. In short, fully phonetic balance database can ensure the output speech quality while small footprint enables it to be implemented into embedded system.

Reference

[1] J. Zhao, *et al.*, "Pronouncing Rehabilitation of Hearing-Impaired Children Based on Chinese 3D Visual-Speech Database," *Frontier of Computer Science and Technology (FCST), 2010 Fifth International Conference,* pp. 625-630, 2010.

[2] A. Chalamandaris, *et al.*, "A unit selection text-to-speech synthesis system optimized for use with screen readers," *Consumer Electronics, IEEE Transactions on,* vol. 56, pp. 1890-1897, 2010.

[3] C. A. Martin, *et al.*, "Speech synthesis for people with a visual impairment in digital television receivers," in *Consumer Electronics (ICCE), 2012 IEEE International Conference on,* 2012, pp. 546-547.

[4] S. Karabetsos, *et al.*, "Embedded unit selection text-to-speech synthesis for mobile devices," *Consumer Electronics, IEEE Transactions on,* vol. 55, pp. 613-621, 2009.

[5] K. Sang-Jin, *et al.*, "HMM-based Korean speech synthesis system for hand-held devices," *Consumer Electronics, IEEE Transactions on,* vol. 52, pp. 1384-1390, 2006.

[6] Y. Ishikawa, *et al.*, "Speech synthesis method based on application-specific synthesis units and its implementation on a 32-bit microprocessor," *Consumer Electronics, IEEE Transactions on,* vol. 45, pp. 980-985, 1999.

[7] H. Zen, *et al.*, "Statistical parametric speech synthesis," *Speech Communication,* vol. 51, pp. 1039-1064, 2009.

[8] T. Tian-Swee and Sh-Hussain, "Corpus design for Malay Corpus-based Speech Synthesis System," *American Journal of Applied Sciences,* vol. 6, pp. 696-702, 2009.

[9] T. Chee-Ming, *et al.*, "Automatic phonetic segmentation of Malay speech database," in *Information, Communications & Signal Processing, 2007 6th International Conference on*, 2007, pp. 1-4.

[10] B. Milner, "Speech Feature Extraction and Reconstruction," in *Automatic Speech Recognition on Mobile Devices and over Communication Networks*, ed: Springer London, 2008, pp. 107-130.

[11] A. P. Dempster, *et al.*, "Maximum Likelihood from Incomplete Data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 39, pp. 1-38, 1977.

[12] S. J. Young, *et al.*, "Tree-based state Tying for High Accuracy Acoustic Modelling," presented at the In: Proceedings of ARPA Human Language Technology Workshop, 1994.

[13] T. Yoshimura, *et al.*, "Duration Modeling for HMM-based Speech Synthesis," in *Proc. of ICSLP*, 1998, pp. 29-32.

[14] K. Tokuda, *et al.*, "Speech Parameter Generation Algorithms for HMM-based Speech Synthesis," *In: International Conference on Acoustics, Speech and Signal Processing (ICASSP 2000),* pp. 1315-1318, 2000.

[15] H. Kawahara, *et al.*, "Reconstructing speech representtions using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication,* vol. 27, pp. 187-207, 1999.

[16] H. Kawahara, *et al.*, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *In: 2nd MAVEBA*, 2001.

[17] N. Seman and K. Jusoff, "Automatic Segmentation and Labeling for Spontaneous Standard Malay Speech Recognition," in *Advanced Computer Theory and Engineering, 2008. ICACTE '08. International Conference on,* 2008, pp. 59-63.