Journal of Computer Science 10 (2): 316-324, 2014
ISSN: 1549-3636
© 2014 Science Publications
doi:10.3844/jcssp.2014.316.324 Published Online 10 (2) 2014 (http://www.thescipub.com/jcs.toc)

LOW FOOTPRINT HIGH INTELLIGIBILITY MALAY SPEECH SYNTHESIZER BASED ON STATISTICAL DATA

Lau Chee Yong and Tan Tian Swee

Medical Implant Technology Group (MediTEG),
Cardiovascular Engineering Center, Material Manufacturing Research Alliance (MMRA),
Faculty of Biosciences and Medical Engineering, Universiti Teknologi Malaysia, Malaysia

Received 2013-07-08; Revised 2013-08-13; Accepted 2013-11-19

ABSTRACT

Speech synthesis plays a pivotal role nowadays. It can be found in various daily applications such as in mobile phones, navigation systems, languages learning software and so on. In this study, a Malay language speech synthesizer was designed using hidden Markov model to improve the performance of current Malay speech synthesizer and also extend Malay speech technology. Statistical parametric method was utilized in this study. The database was constructed to be balanced with all the phonetic sample appeared in Malay language. The results were rated by 48 listeners and obtained a moderate high rating ranging from 3.79 to 4.23 out of 5. The computed Word Error Rate is 7.1%. The total file size is less than 2 Megabytes which means it is suitable to be embedded into daily application. In conclusion, a Malay language speech synthesizer was designed using statistical parametric method with hidden Markov model. The output speech was verified to be good in quality. The file size is small indicates the feasibility to be used in embedded system.

Keywords: Speech Synthesis, Hidden Markov Model, Phonetic Balanced, Footprint

1. INTRODUCTION

Speech is the medium of communication between people. The goal of speech synthesis is to generate speeches from text using computer (Rashad *et al*., 2010). There is a difference between any talking machine and speech synthesizer. Talking machine only playback what is already preset like cassette player while speech synthesizer is used to produce unprecedented utterances (El-Bakry *et al*., 2011). Speech synthesis technique is importantin our daily life. It is incorporated in various systems such as car navigation, screen reader, language learning, telecommunication and so on. To embed the speech synthesizer into these devices, smaller engine size is more preferred.

Currently, only unit selection method has been utilized in designing a Malay language speech synthesizer (Tan, 2008). It includes the recorded speech database into the engine. However, this is not practical to be embedded into daily appliance (Gros, 2006) because of its size. So in this study, we built a Malay language speech synthesizer based on statistical parametric method (Zen *et al*., 2009) using Hidden Markov Model (HMM). We hope to extend the speech technology in Malay language by applying a newer technique and design a smaller size of speech synthesis engine to suit the demand of various devices. We also hope to go beyond the limited domain usage of Malay speech synthesizer by creating a speech synthesizer which can be used at various purposes (Tiun *et al*., 2012).

1.1. Literature Review

Nowadays, two speech syntheses have gained a significant degree of popularity: Unit selection method and statistical parametric method. Both methods have their own merits and each in its own way has made an important contribution. Comparisons between the aforementioned methods are explained in the following paragraph.

Corresponding Author: Tan Tian Swee, Medical Implant Technology Group (MediTEG), Cardiovascular Engineering Center, Material Manufacturing Research Alliance (MMRA), Faculty of Biosciences and Medical Engineering, Universiti Teknologi Malaysia, Malaysia

1.2. Comparison of State-of-the-Art Speech Synthesis System

Corpus based speech synthesis for Malay language has been constructed by Tan (Swee and Salleh, 2008) utilizing unit selection method (Hunt and Black, 1996; Black and Campbell, 1995). That requires a large database consists of recorded speech using real human voice. Subsequently, required utterance is then collected from database and converted into speech. For optimum performance, unit selection method should possess minimum path searching for a sequence of speech units (Lim *et al*., 2012). Therefore, this method involves less signal processing or no signal processing. Unit selection method is popular attributable to its high intelligibility and naturalness of output speech (Alias *et al*., 2011). However, demands larger database for better quality (Barra-Chicote *et al*., 2010). Furthermore, this method is not flexible enough to pronounce unknown word in database (Zen *et al*., 2009). Despite its relatively higher quality on overall resultant synthesized speech, the performance is inconsistent along the whole sentence (Toda *et al*., 2004). Statistical parametric speech synthesis has been introduced (Zen *et al*., 2009) as an alternative to unit selection method. This system simultaneously models spectrum, excitation and duration of speech using mathematical model such as Hidden Markov Model (HMM) to generate speech waveform from the model (Zen *et al*., 2007). Statistical parametric speech synthesis system (Zen *et al*., 2009) has been widely studied recently. Its flexibility in generating speech in other languages like Korean (Sang-Jin *et al*., 2006) and Romanian (Stan *et al*., 2011) becomes the preference of most researchers. Moreover, the quality of speech generated from this method has reached a satisfactory level if compared with unit selection method (Karaiskos *et al*., 2008). There are two major advantages of using statistical parametric speech synthesis vis-à-vis unit selection method: Its capability to model unknown word omitted from the database and smaller required database size because statistical parametric method usually stores statistics of acoustic models rather than multi templates of speech units. This method is more reliable and robust to noise and fluctuation of voice because this method utilized the averaging of statistic in generating speech which means the feature of synthesized speech is normalized. So, this is immune to noise and voice fluctuation occurred during recording session. Unlike unit selection method, tuning parameters in statistical parametric speech synthesis method are fewer than in unit selection method. Moreover, the parameters can be separately controlled so the covariance effect is lesser. Other advantages of using statistical parametric speech synthesis method are easy to extend to other language, smaller runtime engine and easy to change the output voice characteristic by using adaptation method (Tokuda *et al*., 2002).

Another problem arises when constructing speech synthesis system which is regarding the database. By reviewing works done by other researchers (Stan *et al*., 2011), database was created by randomly collect the words from articles, story books, newspapers and online resources. This might possess risks because not all diphonewould exist in the database. Severaldiphonesmight found inexistence in database and the missing diphones were then not trained. Moreover, due to randomly collect the words, certaindiphones occurrence in database might less than 3 times. This shown that random collection of the words from any source is possibly loses some diphones and end up with larger size of incomplete database. In order to acquire good quality of synthetic speech, it is important to have a phonetically balanced database in constructing a competitive speech synthesizer. A database complete with all the phonetic samples is an essential key point to increase the robustness of speech synthesizer regardless of the method used (unit selection and statistical parametric method).

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.

1.3. Introduction of Malay Language

Malay language or Bahasa Melayu is the official language in Malaysia, Indonesia and Brunei. This language possesses agglutinative characteristic where the words in Malay language are formed by joining syllables. “Standard Malay” (SM) is a term that describe the style of Malay language to be the norm of that language. Regional characteristic in a country like Malaysia also affects the language to different dialects. Beside standard Malay, there are three dialects namely Kelantan Malay, Ulu Muar Malay and Langkawi Malay (Seman and Jusoff, 2008) andall of them are different in certain accent and pronunciation. Basically, Malay language has its own special characteristics. First, Malay language is a type of phonetic language and written in Roman characters. Second, the syllables in Malay language are pronounced almost equally and it is not a tonal language. Generally, six (6) vowels and 29 consonants can be found in Malay language. The basic syllable structure of Malay language is produced according to three syllabification rule. Malay was defined as a Type III language by linguists; consist of

Consonant-Vowel (CV) and Consonant-Vowel-Consonant (CVC). These combinations are most commonly found in every Malay primary word (Teoh, 1994).

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. 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 inpronunciation (Chee-Ming *et al*., 2007). **Table 1** lists down the phonemes according to its group.

2. MATERIALS AND METHODS

2.1. Text Corpus

The words of the lexicon database were gathered from online Malay news such as BeritaHarian, Bernama, Cybersing, Malaysia Kini, Utusan and primary school.

Malay Language textbooks to collect more frequently used words in daily life. Total of 10 million words were 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 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.

It is not practical to compose the sentence script using all of the words found from the resources. Hence,a threshold of 70% of highest frequency words was chosen as a benchmark for the words selection. Total of 1451 words found in that range of most frequently used words. The reason for choosing 70% is because the number of words collected is practical for training and covers a sufficient percentage of words. **Table 2** shows the number of words in different percentage of frequency of appearance. In the word collection, top 10 of highest frequency words were computed. 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 3** 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.

We hired a Malay native female speaker to record the database to ensure correct pronunciation and natural speech. Standard Malay was used as the intonation and no other accent was used thoughout the recording session. The recording session was conducted in a quiet room using SM48 Cardioid Dynamic Vocal Microphone from Shure Company. Surrounding noise was measured before every recording session and only proceeded if the noise is low. The speaker spoke for a few minutes before each recording session to adapt a natural speaking style and intonation. The recording is done in 96kHz of sampling frequency and the recorded speeches were downsampled into 16kHz. It is because 16kHz frequency is sufficient to acquire good training output and a practical processing time.

2.2. Core Structure of Speech Synthesis Engine

Figure 1 shows the core structure of a speech synthesizer (Chomphan, 2011). We adopted statistical parametric speech synthesis method introduced by (Zen *et al*., 2009). Excitation and spectral parameters were extracted from recorded speech database and trained into HMM with iterative training. To estimate the model parameter, Maximum Likelihood criterion (ML) is used according to the Equation (1) below:

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

Where:

λ = is a set of model parameters

O = is a set of training data

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 Equation (2):

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

After that, a speech waveform is rendered from the parameter using a vocoder. Three important elements for statistical parametric speech synthesis are mel-cepstrum, generalized log f0 and band limited aperiodicity measure.

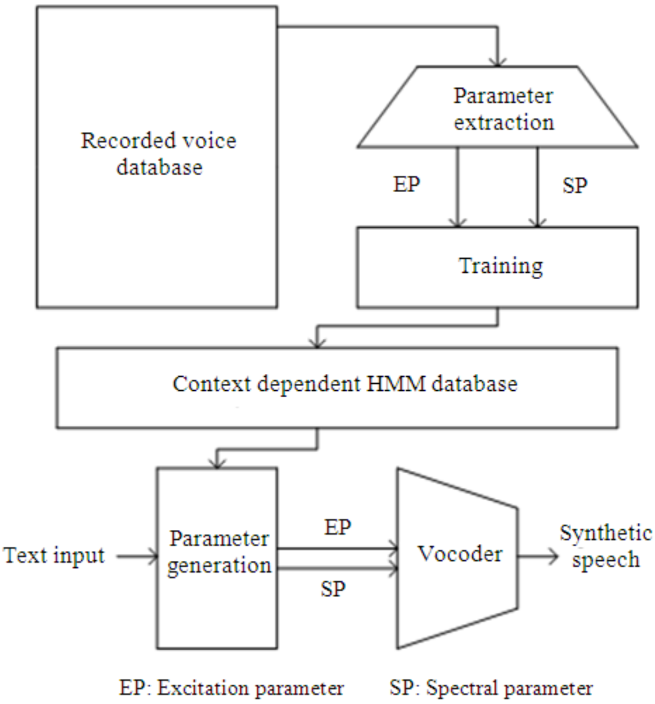

Fig. 1. Core structure of speech synthesizer

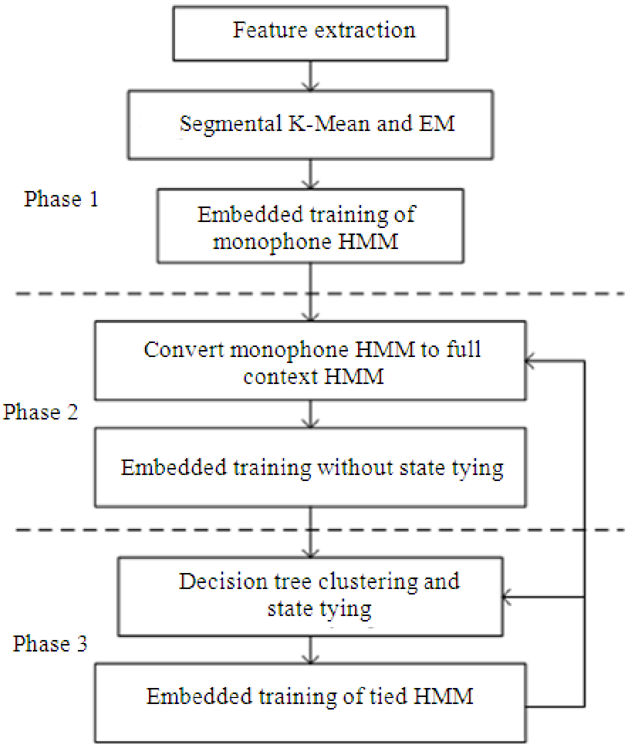
Feature extraction
Segmental K-Mean and EM
Phase 1
Embedded training of monophone HMM
Convert monophone HMM to full context HMM
Phase 2
Embedded training without state tying
Decision tree clustering and state tying
Phase 3
Embedded training of tied HMM

Fig. 2. Block diagram of training process

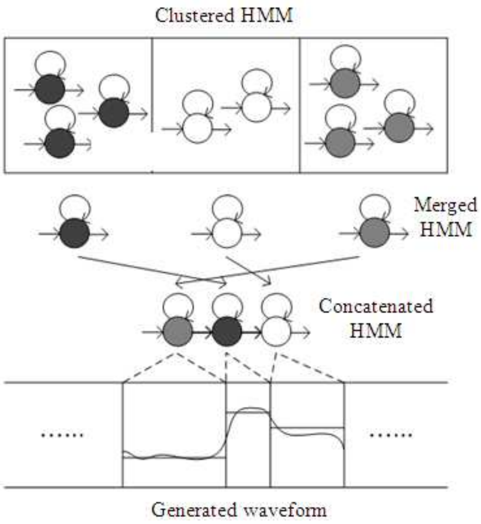

Fig. 3. Block diagram of speech synthesis process

Table 1. 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/

Table 2. Word coverage according to frequency of occurrence

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

Table 3. Top 10 Malay words with highest frequency of 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

The training process generates the average of these parameters and the output speech is synthesized from the normalized data.

The overall training process is in three phases as in **Fig. 2**. At the first phase, the monophone HMMs are trained with the initial segments of database script and speech using segmental k-means to classify the group of phonemes and Expectation-Maximization (EM) algorithm (Dempster *et al*., 1977) was used to perform embeddedtraining 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 (Young *et al*., 1994) 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 utterance was converted into context-labeled sequence using a text-processor. Then, the corresponding context-dependent HMMs were gathered and concatenatedinto same manner as the target utterance. The state durations of the HMM were determined so as to maximize the probability of state durations (Yoshimura *et al*., 1998). Speech parameter generation algorithm like in Case 1 of (Tokuda *et al*., 2000) was used to determine the mel-cepstral coefficients and log F0 values. By the inclusion of dynamic coefficients, the output speech is smoothen and constrained to be realistic. Finally, STRAIGHT mel-cepstralvocoderwas used to render the speech waveform (Kawahara *et al*., 1999; 2001). The process can be illustrated as in **Fig. 3**.

3. RESULTS

We hired 48 listeners to conduct a listening test to rate the synthetic speech based on their opinion of:

- Clarity. Listeners were asked to rate the clarity of synthetic speech using a 5-point scale from 1 (Totally not clear) to 5 (Very clear). Higher rate implies listeners can understand the synthetic speech without difficulties
- Naturalness. Listeners used a 5-point scale from 1 (Totally not natural) to 5 (Very natural) to rate the naturalness of synthetic speech based on their opinion regardless of intelligibility of the speech
- Similarity. Listeners rated the synthetic speech with a 5-point scale from 1 (Totally not similar) to 5 (Very similar) based on their opinion of similarity between output speech and original speech
- Intelligibility. Listeners were asked to transcribe the synthetic speech into text and Word Error Rate (WER) was computed to mark the intelligibility level

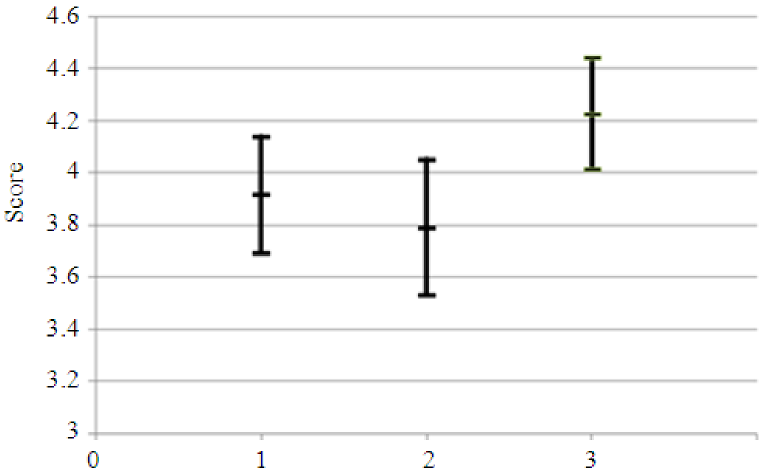

Fig. 4. Graph of listening test result

Table 4. Mean scores and word error rate of listening test

	Mean score with 95% of confidence level
Clarity	3.92±0.225
Naturalness	3.79±0.261
Similarity	4.23±0.212
	WER
Intelligibility	7.10%

Table 5. Footprint of speech synthesize

	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

Each listener was listened to 75 synthetic speeches with headphone to rate clarity, naturalness and similarity in a quiet room. After that, they were asked to transcribe 5 synthetic speeches into text. The formula of Word Error Rate (WER) for intelligibility test is shown below Equation (3):

$$\text{WER} = \frac{S+D+I}{S+D+C} \quad (3)$$

where, S is the number of substitution, D is the number of deletion, I is the number of insertion and C is the number of correct words. The total number of each element was gathered from the transcription of listeners and the WER was computed. Spelling error is not considered as substitution. Thedetail result of listening test is shown in **Table 4 and Fig. 4** below.

From the **Fig. 4**, 1 at x-axis represent clarity while 2 and 3 represent naturalness and similarity respectively.

From the listening test, the variation of the test result is not obvious because of the averaging effect at training stage. Similarity possesses highest rating with 4.23 out of 5 because 39th order of mel-cepstral is being used. In short, this synthesizer performs at a satisfactory level.

The total footprint of this speech synthesizer is less than 2Mbytes with no compression. So it is suitable to become an embedded system in devices (Zen *et al*., 2009; Sang-Jin *et al*., 2006). The summary of footprint of the synthesizer is shown in **Table 5**.

4. DISCUSSION

The Malay language characteristics possess as advantages in this study. The letter to sound rule of Malay language is straight forward and almost equally and constantly. Moreover, it is not a tonal language. Tonal language requires extra implementation in order to synthesize better quality of speech (Chomphan, 2010). Besides, Malay language is written in Roman character which means the grapheme to phoneme conversion is also excluded in speech synthesis process (Seman and Jusoff, 2008). In short, fully phonetic balance database can ensure the output speech quality while small footprint enables it to be implemented into embedded system.

Generally, statistical parametric speech synthesizer is significant to current technology nowadays. Its low footprint and system size is preferable to be embedded into various applications. This characteristic also enables it to be portable easily. Besides, with only limited of training data, it can generate fairly acceptable quality of out-of-corpus word. This feature is friendly to different field which uses synthetic speeches. Moreover, this system is robust to noise as it generates average features from database to synthesize speech. Unlike unit selection method which uses concatenation of real natural speech to generate synthetic speech, this method is not much affected by the quality fluctuation of training database.

However, if compared to unit selection method, the naturalness of synthetic speech using this method is lower than using unit selection method. It is because the synthetic speech is generated using a wave synthesizer while unit selection uses real speech segments to compose a speech. But the overall quality of synthetic speech in statistical method is acceptable.

5. CONCLUSION

We proposed a newer technique of speech synthesis in Malay language to extend the speech technology in this language. By comparing with state to the art unit selection technique, this method offers some advantages includes 95.1% reduction in total file size to make it more practical to be embedded in daily used devices. We designed the lexicon database carefully by taking care of all the phonemes in Malay language and composed phonetic balanced scripts. The output synthetic speech was rated by 48 listeners with opinion scores and Word Error Rate (WER). The synthetic speech possesses a satisfactory level of quality ranging from 3.79 to 4.23 out of 5 and the WER is 7.1%. In short, phonetic balanced database could provide a good synthetic speech and smaller file size is preferred to be embedded into daily used devices.

6. FUTURE WORK

The process of preparing training database is expensive. It involves composition of scripts and recording of speeches. In this study, this process has taken approximately 2 months from script creation until the speech was recorded. However, there are some readily available online resources of recorded speech and text transcription such as audio book, podcast, broadcast and so on which provide high quality recorded speeches. This resource can be useful if the clear segments of speeches are extracted to be the training database. To get a suitable online speech and text, some investigation must be done to ensure the resources are complete with all phonetic samples and the topic of speech should not be biased into certain topics. So a richer phonetic context scripts can be obtained. The method of collect and refine readily online data could be an interesting research topic in future.

7. ACKNOWLEDGEMENT

The researchers would like to thank IJN for their professional opinions and involvement, Ministry of Higher Education (MOHE), Universiti Teknologi Malaysia (UTM) and UTM Research Management Centre (RMC) for supporting this research project under grant code Q.J130000.2445.00G70.

8. REFERENCES

Alias, F., L. Formiga and X. Llora, 2011. Efficient and reliable perceptual weight tuning for unit-selection text-to-speech synthesis based on active interactive genetic algorithms: A proof-of-concept. Speech Commun., 53: 786-800. DOI: 10.1016/j.specom.2011.01.004

Barra-Chicote, R., J. Yamagishi, S. King, J.M. Montero and J. Macias-Guarasa, 2010. Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. Speech Commun., 52: 394-404. DOI: 10.1016/j.specom.2009.12.007

Black, A.W. and N. Campbell, 1995. Optimising selection of units from speech databases for concatenative synthesis. Edinburgh Res. Arch., 1: 581-584.

Chee-Ming, T., S.H. Salleh, T. Tian-Swee and A.K. Ariff, 2007. Automatic phonetic segmentation of Malay speech database. Proceedings of the 6th International Conference on Information, Communications and Signal Processing, Dec. 10-13, Singapore, pp: 1-4. DOI: 10.1109/ICICS.2007.4449574

Chomphan, S., 2010. Tone question of tree based context clustering for hidden Markov model based Thai speech synthesis. J. Comput. Sci., 6: 1474-1478. DOI: 10.3844/jcssp.2010.1474.1478

Chomphan, S., 2011. Analysis of decision trees in context clustering of hidden Markov model based Thai speech synthesis. J. Comput. Sci., 7: 359-365. DOI: 10.3844/jcssp.2011.359.365

Dempster, A.P., N.M. Laird and D.B. Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc., 39: 1-38. DOI: 10.2307/2984875

El-Bakry, H.M., M.Z. Rashad and I.R. Isma'il, 2011. Diphone-based concatenative speech synthesis systems for Arabic language. Proceedings of the 10th WSEAS International Conference on Circuits, Systems, Electronics, Control and Signal Processing, (CSP' 11), World Scientific and Engineering Academy and Society (WSEAS) Stevens Point, Wisconsin, USA., pp: 81-86.

Gros, J.Z., 2006. Text-to-speech synthesis for embedded speech communicators. Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, (EDB' 06), World Scientific and Engineering Academy and Society (WSEAS) Stevens Point, Wisconsin, USA., pp: 189-193.

Hunt, A.J. and A.W. Black, 1996. Unit selection in a concatenative speech synthesis system using a large speech database. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 7-10, IEEE Xplore Press, Atlanta, GA., pp: 373-376. DOI: 10.1109/ICASSP.1996.541110

Karaiskos, V., S. King, R.A.J. Clark and C. Mayo, 2008. The blizzard challenge 2008. University of Edinburgh.

Kawahara, H., I. Masuda-Katsuse and A. De Cheveigne, 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Commun., 27: 187-207. DOI: 10.1016/S0167-6393(98)00085-5

Kawahara, H., J. Estill and O. Fujimura, 2001. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. Proceedings of the 2nd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, Sept. 13-15, Firenze, Italy, pp: 59-64.

Lim, Y.C., T.S. Tan, S.H.S. Salleh and D.K. Ling, 2012. Application of genetic algorithm in unit selection for Malay speech synthesis system. Expert Syst. Applic., 39: 5376-5383. DOI: 10.1016/j.eswa.2011.11.047

Rashad, M.Z., H.M. El-Bakry, I.R. Isma'il and N. Mastorakis, 2010. An overview of text-to-speech synthesis techniques. Proceedings of the 4th International Conference on Communications and Information Technology, (CIT' 10), World Scientific and Engineering Academy and Society (WSEAS) Stevens Point, Wisconsin, USA., pp: 84-89.

Sang-Jin, K., K. Jong-Jin and H. Minsoo, 2006. HMM-based Korean speech synthesis system for hand-held devices. IEEE Trans. Consumer Electron., 52: 1384-1390. DOI: 10.1109/TCE.2006.273160

Seman, N. and K. Jusoff, 2008. Automatic segmentation and labeling for spontaneous standard Malay speech recognition. Proceedings of the International Conference on Advanced Computer Theory and Engineering, Dec. 20-22, IEEE Xplore Press, Phuket, pp: 59-63. DOI: 10.1109/ICACTE.2008.150

Stan, A., J. Yamagishi, S. King and M. Aylett, 2011. The Romanian Speech Synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate. Speech Commun., 53: 442-450. DOI: 10.1016/j.specom.2010.12.002

Swee, T.T. and S.H.S. Salleh, 2008. Corpus-based Malay text-to-speech synthesis system. Proceedings of the 14th Asia-Pacific Conference on Communications, Oct. 14-16, IEEE Xplore press, Tokyo, pp: 1-5.

Tan, T.S., 2008. Implementation of phonetic context variable length unit selection module for Malay text to speech. J. Comput. Sci., 4: 550-556. DOI: 10.3844/jcssp.2008.550.556

Teoh, B.S., 1994. The Sound System of Malay Revisited. Dewan Bahasa dan Pustaka, Ministry of Education, Malaysia, Kuala Lumpur, ISBN-10: 9836241841. pp: 150

Tiun, S., R. Abdullah and T.E. Kong, 2012. Restricted domain Malay speech synthesizer using syntax-prosody representation. J. Comput. Sci., 8: 1961-1969. DOI: 10.3844/jcssp.2012.1961.1969

Toda, T., H. Kawai and M. Tsuzaki, 2004. Optimizing sub-cost functions for segment selection based on perceptual evaluations in concatenative speech synthesis. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 17-21, IEEE Xplore Press, pp: I-657-660. DOI: 10.1109/ICASSP.2004.1326071

Tokuda, K., T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura, 2000. Speech parameter generation algorithms for HMM-based speech synthesis. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE Xplore Press, Istanbul, Jun. 05-09, pp: 1315-1318. DOI: 10.1109/ICASSP.2000.861820

Tokuda, K., Z. Heiga and A.W. Black, 2002. An HMM-based speech synthesis system applied to English. Proceedings of the IEEE Workshop on Speech Synthesis, Sep. 11-13, IEEE Xplore Press, pp: 227-230. DOI: 10.1109/WSS.2002.1224415

Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, 1998. Duration Modeling for HMM-based speech synthesis. Proceedings of the in ICSLP, pp: 29-32.

Young, S.J., J.J. Odell and P.C. Woodland, 1994. Tree-based state tying for high accuracy acoustic modeling. Proceedings of the Workshop on Human Language Technology, (HLT' 94), Association for Computational Linguistics Stroudsburg, PA, USA., pp: 307-312. DOI: 10.3115/1075812.1075885

Zen, H., K. Tokuda and A.W. Black, 2009. Statistical parametric speech synthesis. Speech Commun., 51: 1039-1064. DOI: 10.1016/j.specom.2009.04.004

Zen, H., T. Nose, J. Yamagishi, S. Sako and T. Masuko *et al*., 2007. The HMM-based speech synthesis system (HTS) version 2.0. Proceedings of the 6th ISCA Workshop on Speech Synthesis, Aug. 22-24, Bonn, Germany, pp: 294-299.