Corpus-based Malay Text-to-Speech Synthesis System

Tan Tian Swee and Sheikh Hussain Shaikh Salleh
Faculty of Biomedical and Health Science Engineering
Universiti Teknologi Malaysia
81310 UTM Skudai
Johor, Malaysia
E-mail: tantianswee@hotmail.com, hussain@fke.utm.my

***Abstract*— The main problem with current Malay text-to-speech (TTS) synthesis system is the poor quality of the generated speech sound. This poor quality is resulted from the inability of traditional TTS system to provide multiple choices of unit for generating more accurate synthesized speech. Most of the current available Malay TTS systems are utilizing diphone concatenation that only support a single unit for each existing diphone, thus it cannot provide more accurate selection of speech unit for concatenation]. This project has implemented a variable length unit selection Malay text to speech system that is capable of providing more natural and accurate unit selection for synthesized speech. This paper proposes a method of combining both linguistic context and feature distance cost for selecting the best match unit. A set of digitized Malay word has been collected from Malay internet news for Malay word frequency count. 381 sentences have been designed which cover around 70 percent of high frequency words from 10 million digitized word obtained from Malay internet news. Then a unit selection method has been implemented to provide the capability of selecting a speech unit not only limited to phoneme, diphone or triphone but also a string of phonemes that can be matched directly to the database. A set of listening test namely Modify Rhythm Test (MRT) has been carried out with 35 participants, which represented 86 percent of accuracy.**

I. INTRODUCTION

Malay TTS system is a speech synthesis tool that is able to pronounce any Malay input raw texts aloud [2][3]. This system is constructed by using the unit selection concatenate synthesis technology. Malay TTS system consists of a few main components, which are Tokenizer, Phonetizer, unit selection engine, and speech synthesis engine. Figure 1 shows the architecture of MTTS.

The first component of TTS system is the tokenizer that pre-processes the input raw texts and converts it into normalize full text [3]. The second component is the phonetizer which produces the target unit sequence for input text. This target sequence unit will then be fitted into the unit selection engine to search for the best match unit. Once the best match unit is found, it will be passed to the speech synthesizer for concatenation, smoothing and modification of the speech [3].

In this project, a set of speech corpus has been designed to minimize the database and provide the best coverage of high frequency words that is capable of producing synthetic speech with high naturalness and minimized database size.

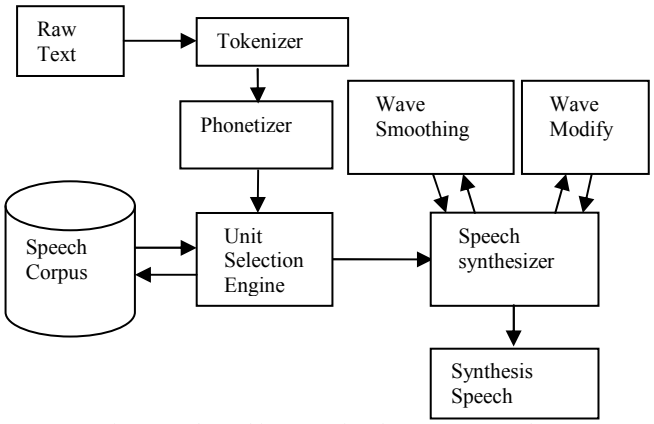

Figure 1: The architecture of Malay Text to Speech

II. SPEECH CORPUS DESIGN

As in Figure 2, the process of corpus building starts from text corpus design and then speech corpus design.

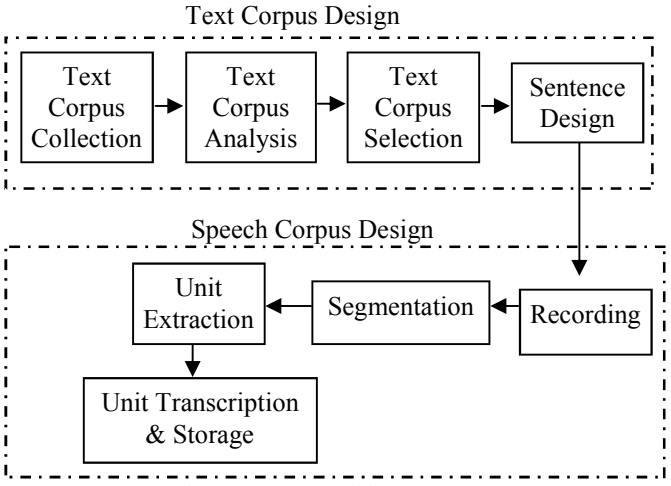

Figure 2: Corpus Building Process.

For text corpus design, a set of online Malay News consisting of 10 million words has been collected in 3 months.

Since all the texts are in html format, a text processing tool has been designed to extract text automatically from html file and do the word frequencies count [5]. The information of 10 millions words is shown in Table 1. The design of a well utilized speech corpus requires determining an optimal set of elements for storage in the database [4]. To help in finding an element set of optimal size and composition along these guidelines, some statistical analyses have been conducted [4].

Total text	10027126
Duration of collection	3 months
Total existing Malay word	115738

Table 1: Detail of 10 million Malay words from internet News

A. Sentence List Design

Table 2 shows the percentage of coverage for high frequency words. Figure 3 shows the coverage graph from the analysis of high frequency words. 1451 high frequency words have covered 70% of 10 million words and it is the most optimum category which is not too high in terms of total words and has sufficient percentage of coverage. Thus, the sentence list for recording can be minimized and is able to support the word based concatenation to be more natural.

Table 3 shows the significance of the top 10 highest frequency words among 10 million words of Malay texts. It shows that the total of 10 highest frequency words has contributed 10.58 percent of the 10 millions words. This has given us the idea on the significance of high frequency word in designing the database.

Category	Total words
60% of high frequency text	747
65% of high frequency word	1025
70% of high frequency word	1451
80% of high frequency word	2592

Table 2: Word coverage for high frequency words.

	Word	counting
1	yang	282200
2	dan	253097
3	untuk	91374
4	tidak	84443
5	pada	60179
6	akan	58222
7	saya	55550
8	kepada	55497
9	mereka	55175
10	ke	42310
	Total :	1038047
	% of 10 millions:	10.57858899

Table 3: 10 highest frequency words

The 1451 words list that cover 70% of the high frequency words have been used to design the sentences. The same list of sentences has been used to create a support for word based concatenation module to provide higher naturalness [6][7][8]. Altogether 381 sentences have been designed. Then, a female voice pronouncing the sentences was recorded and segmented using segmentation tools designed by the Center for BioMedical Engineering, UTM.

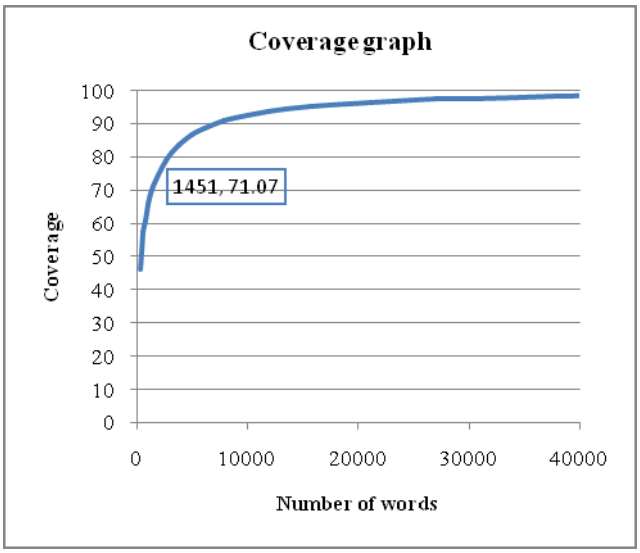

Figure 3: Coverage Graph for 10 Million Words

B. Speech Unit Database

There are three main things in speech unit database, the speech wave, speech unit transcription file and speech unit's speech feature, Mel-Frequency Cepstrum Coefficient (MFCC).

The smallest speech unit, phonemes, were then extracted from the segmented sentences and put into a speech unit database folder as shown in Figure 4. The total unit for each phoneme is listed in Table 4. It can be seen that the phoneme "a" and "e" are the highest phoneme in the list.

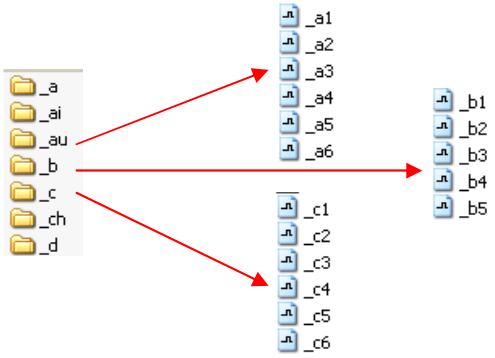

Figure 4: Speech unit database.

The transcription file for speech unit is as shown in Figure 5. It consists of the speech unit in phonetic context and the

correspondent speech wave. This information will be used for the first process of unit selection that will be discussed later.

Pho	Total	%
_a	107	0.64
_ai	4	0.02
_au	1	0.01
_b	256	1.52
_c	29	0.17
_d	269	1.6
_e	3	0.02
_eh	10	0.06
_f	17	0.1
_g	30	0.18
_h	65	0.39
_i	74	0.44
_ia	12	0.07
_j	49	0.29
_k	248	1.47
_kh	8	0.05
_l	72	0.43
_m	447	2.66

Pho	Total	%
_n	33	0.2
_ny	2	0.01
_o	21	0.12
_p	320	1.9
_r	58	0.34
_s	258	1.53
_sy	5	0.03
_t	178	1.06
_u	42	0.25
_v	5	0.03
_w	18	0.11
_y	59	0.35
_z	1	0.01
a	3076	18.3
ai	97	0.58
au	26	0.15
b	253	1.5
c	77	0.46

Pho	Total	%
d	313	1.86
e	1448	8.6
eh	124	0.74
f	38	0.23
g	169	1
h	374	2.22
i	970	5.76
ia	87	0.52
io	3	0.02
iu	1	0.01
j	164	0.97
k	665	3.95
kh	7	0.04
l	514	3.05
m	492	2.92
n	1293	7.68
ng	500	2.97
ny	91	0.54

Pho	Total	%
o	206	1.22
p	276	1.64
q	1	0.01
r	838	4.98
s	410	2.44
sy	8	0.05
t	652	3.87
u	696	4.13
ua	107	0.64
ui	2	0.01
v	10	0.06
w	72	0.43
y	52	0.31
z	13	0.08

Table 4: Total units after extract the phoneme units from the carrier sentences.

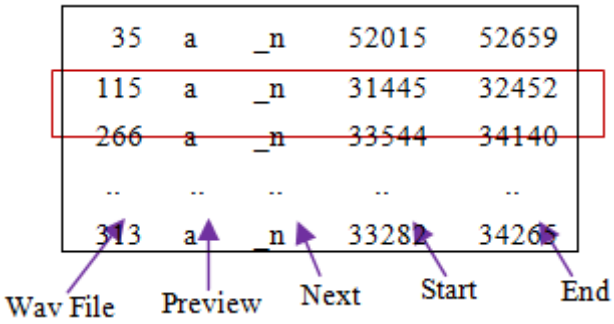

Figure 5: Phonetic transcription for speech unit.

The speech features, MFCC for the first and last frame of each speech unit as shown in Figure 6 were then stored in the database for unit searching and matching. It was used for distance measurement to match two minimized distance speech unit when more than 2 speech units fulfill the phonetic context.

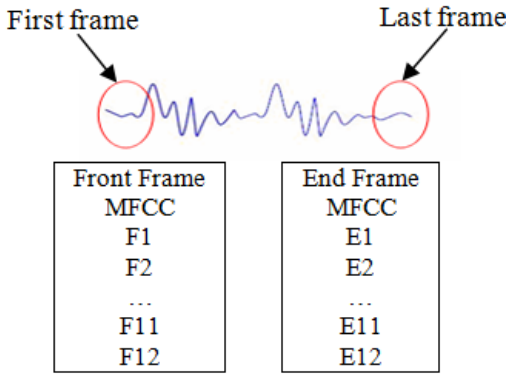

Figure 6: First and last frame speech feature.

MFCCs [10] are representative of the real cepstrum for a windowed short time signal derived from the fast Fourier Transform (FFT) of the speech signal. Its difference with the real cepstrum is that a nonlinear, perceptual motivated frequency scale is being used, which approximates the behavior of the human auditory system. This feature has been used for the later part of selecting unit using the spectral distance method that will be discussed later.

III. CONCATENATION ENGINE

Concatenation Engine for MTTS as shown in Figure 6 will first search the biggest speech unit from the word database which consists of 70 percent of high frequency words. If the word exists in the list, it will then apply the words directly. If the word cannot be found in the word database, it will go through the unit selection engine.

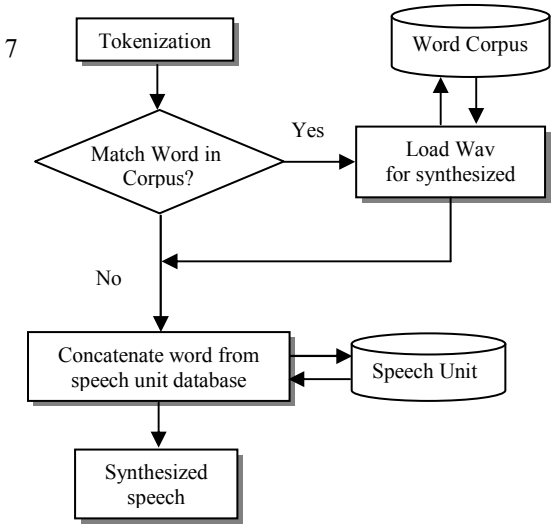

Figure 7: Speech Concatenation Engine.

IV. CONCATENATION ENGINE

The unit selection engine consists of two main modules as shown in Figure 7. The first module is the phonetic context searching module which selects the best match of unit sequence with the phonetic context. Then, if the best match of unit exists more than once, it will use the second module to find the minimized spectral distance speech unit as the best match unit.

A. Phonetic Context Preprocessing

For concatenating non-existing word, the target sequence of speech unit in phoneme level will be generated. An example of target sequence of speech unit in phonetic context is shown in Figure 8. This target sequence will first pass through the linguistic context searching module. This module will match the phoneme with its previous and current phoneme context. If the best match target unit is more than one, then it will pass to the feature distance measuring module to find the minimized spectral distance. This engine is not limited to

phoneme or diphone or triphone but a variable length of unit selection.

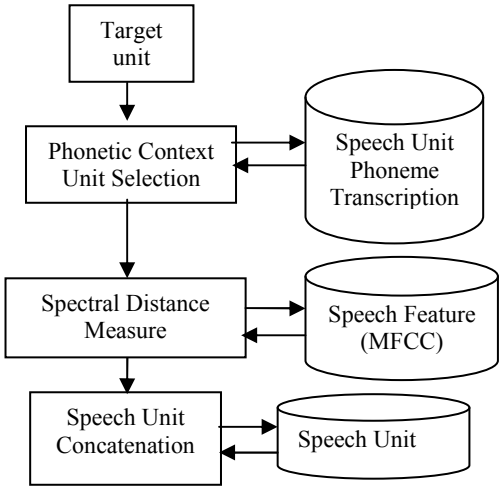

Figure 8: Unit Selection Engine.

B. *Spectral Distance Measure*

The spectral representation is often a smoothed spectral envelop, possibly as a transformed set of coefficients, which is derived from a short term (frame based) analysis of the speech signal [10]. The speech feature or parameterizations in-use is Mel Frequency cepstral coefficients (MFCC). To measure the difference between two vectors of this speech feature, it needs a distance measurement. The distance measurement used in this case study is Euclidean Distance.

Euclidean distance is the ordinary distance between the two points as shown in Figure 9. The Euclidean distance between two points P = (P1, P2,….., Pn) and Q = (Q1, Q2,….., Qn) in Euclidean n-space, is defined as:

$$\sqrt{(P1-q1)^2+(P2-q2)^2+\cdots+(Pn-qn)^2}=\sqrt{\sum_{i=1}^{n}(Pi-qi)^2} \quad (2.2)$$

This distance is easy to be computed. However, the Euclidean distance does not take into any account of various or covarious of the feature vectors distribution [11][12].

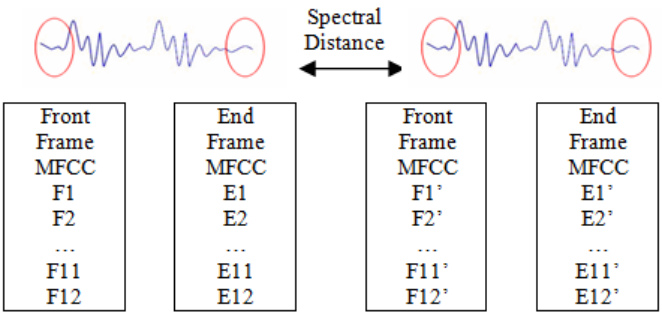

Figure 9: Spectral Distance Measure

C. *Unit Matching example*

The list of units (if it exists more than once after phonetic context matching) will pass through the distance measuring engine to find the minimized spectral distance unit.

It will match the current unit's last frame's MFCC (speech feature), of current unit which is 12 coefficients for this case, with first frame's MFCC of the next unit. Then the next unit that provide the minimized distances will be selected as the most accurate unit and it will be used to match with the following unit until the final unit.

Figure 10 shows an example of an input word "saya suka makan nasi" which means "I like to eat rice" in Malay language. The text will be first tokenize and then phonetize to generate sequence of phoneme. Then, this target sequence will go through the unit selection engine. The unit selection engine will first using phonetic context of target unit to select all best match unit from speech unit database. If more than two unit matches for each target, then it will use spectral distance measure to choose the best match. Then the result of phoneme list and its original sentence and start and end point will be used to concatenate the speech.

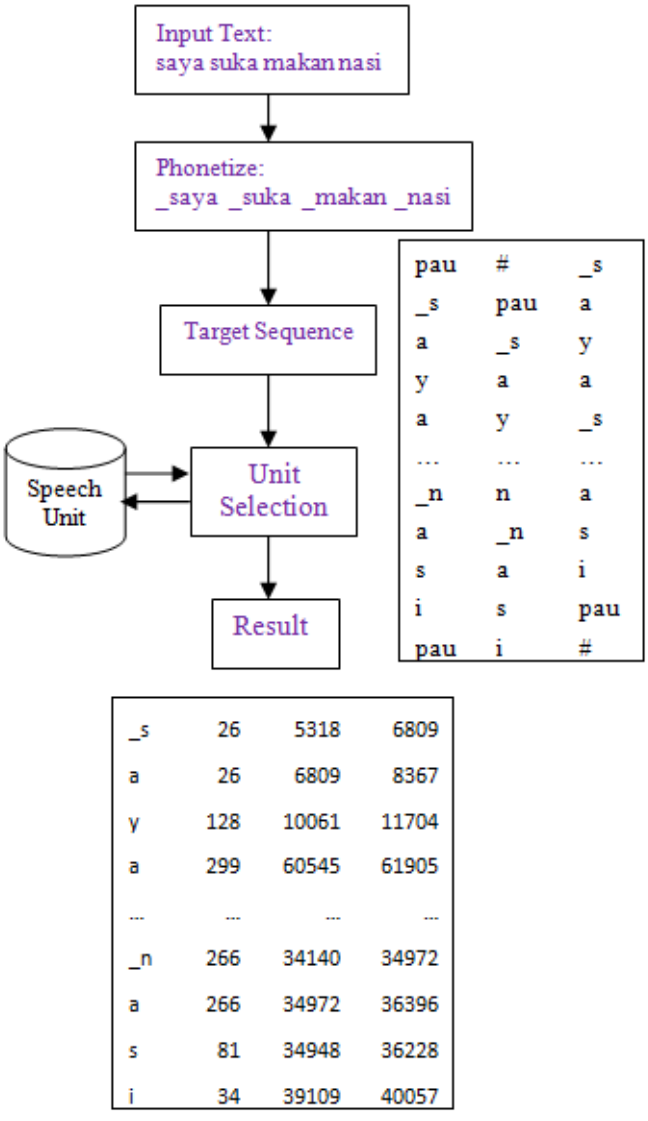

Figure 10: The processing of synthesizing input text using Malay Text to Speech Synthesis system.

V. LISTENING TEST

To evaluate the synthesized speech samples, a set of Modify Rhythm Test Questionnaire, consists of 50 questions, has been set and tested among 35 students. The listeners were instructed to play the test files two times and then guess and select the answer they thought was the best match for the wave sound. The modify rhythm test is shown as Figure 11.

SECTION 2- Modify Rhythm Test (MRT)

For each of the speech samples played twice, choose one which you think is most appropriate for how the speaker sounds.

Question #	Correct Answer	Possible Answers				
		(a)	(b)	(c)	(d)	(e)
1		baja	kaca	baca	bata	baka
2		baju	balu	bayu	baku	baru
3		benang	berang	bedan	beram	belang
4		bidan	bidang	bilang	bijan	bingal
5		bocor	bedoh	bomoh	bolong	borong
6		botol	bokoh	bokop	boroh	botor
7		budur	buduk	buhuk	buhul	bujuk
8		bulur	buluh	bunuh	bunut	buruh
9		busur	butut	butul	busut	buku
10		burung	bunun	buyung	buyur	buyut
11		cabar	calar	cagar	cakar	cadar

Figure 11: Modify Rhythm Test.

VI. TESTING RESULT AND DISCUSSION

The overall score of correct answers for the 35 participants in MRT test is 86.66 percent. The main interface of the system is shown in Figure 12. This system can support for duration, volume and pitch modify that can be used for further study in intonation and expression speech synthesis in Malay Text to Speech Synthesis system.

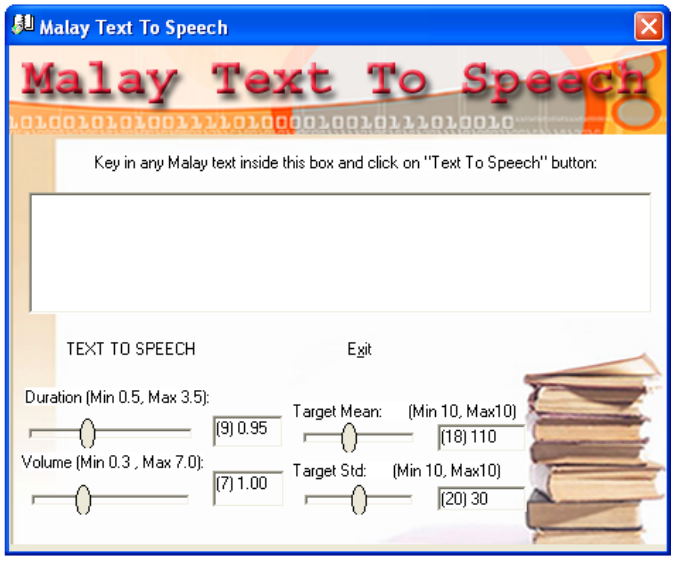

Figure 12: The Malay Text to Speech Synthesis

ACKNOWLEDGMENT

This research project is supported by CBE (Center of Biomedical Engineering) at Universiti Teknologi Malaysia and funded by Ministry of Science, Technology and Innovation (MOSTI), Malaysia under grant "To Develop a Malay Speech Synthesis System for a Standard Platform Compatibility and Speech Compression" Vot 79190.

REFERENCES

[1] Tan, T. S., Sheikh, H. and Aini, H. Building Malay Diphone Database for Malay Text to Speech synthesis System Using Festival Speech Synthesis System. *Proc of The International Conference on Robotics, Vision, information and Signal Processing 2003*. January 22-24. Penang, Malaysia: ROVISP03, 634-648.

[2] Tan, T. S. and Sheikh H. Building Malay TTS Using Festival Speech Synthesis System. *Conference of The Malaysia Science and Technology*, September 2-3. Johor Bahru, Malaysia: MSTC 2002, 120-128.

[3] Tan Tian Swee. *The Design and Verification of Malay Text To Speech Synthesis System.* Master Thesis. Univeristy Technology Malaysia, 2003.

[4] Nagy, A., Pesti, P., Németh, G. and Bőhm, T. (2005). Design Issues of a Corpus-Based Speech Synthesizer. *Hungarian Journal on Communications*, 2005/6, special issue, Budapest, Hungary: pp. 18-24.

[5] Galicia-Haro, S.N. (2003). Using electronic texts for an annotated corpus building. *Fourth Mexican International Conference on Computer Science*. Sept 8-12. Tlaxcala, Mexico: ENC'03, 26-32.

[6] Gaura, P. (2003). Czech speech synthesizer Popokatepetl based on word corpus. *4th EURASIP Conference focused on Video/Image Processing and Multimedia Communications*. July 2-5. Zagreb, Croatia: EC-VIP-MC 2003, Vol. 2, 673–678.

[7] Stöber, K., Portele, T., Wagner, P. and Hess, W. (1999). Synthesis by word concatenation. *Sixth European Conference on Speech Communication and Technology (EUROSPEECH'99)*. September 5-9. Budapest, Hungary: EUROSPEECH'99, Vol. 2, 619-622.

[8] Lewis, E. and Tatham, M. (1999). Word and syllable concatenation in text-to-speech synthesis. *Sixth European Conference on Speech Communications and Technology*, September 5-9. Budapest, Hungary: EUROSPEECH'99, Vol. 2, 615--618.

[9] Rutten, P., Coorman, G., Fackrell, J. and Van Coile, B. (2000). Issues in corpus based speech synthesis. *IEE Seminar on State of the Art in Speech Synthesis* (Ref. No. 2000/058), April 13. Savoy Place, London: 16/1-16/7

[10] Wei HAN, Cheong-Fat CHAN, Chiu-Sing CHOT and Kong-Pang PUN. *An Efficient MFCC Extraction Method in Speech Recognition.* The Chinese University of Hong Kong, 2006.

[11] Yi, J.R.W. (2003). *Corpus-Based Unit Selection for Natural-Sounding Speech Synthesis*. Massachusetts Institute of Technology: Ph.D. Thesis.

[12] Subjective evaluation of join cost and smoothing methods for unit selection speech synthesis Vepa, J.; King, S.; Audio, Speech and Language Processing, IEEE Transactions on Volume 14, Issue 5, Sept. 2006 Page(s):1763 – 1771

[13] New objective distance measures for spectral discontinuities in concatenative speech synthesis Vepa, J.; King, S.; Taylor, P.; Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on 11-13 Sept. 2002 Page(s):223 - 226