Journal of Computer Science 4 (7): 550-556, 2008
ISSN 1549-3636


# Implementation of Phonetic Context Variable Length Unit Selection Module for Malay Text to Speech

Tian-Swee Tan and Sh-Hussain
Faculty of Biomedical Engineering and Health Science, P11, Center for Biomedical Engineering, University Teknologi Malaysia, 81310 UTM Skudai, Johor DT, Malaysia

**Abstract: Problem statement:** The main problem with current Malay Text-To-Speech (MTTS) synthesis system is the poor quality of the generated speech sound due to the inability of traditional TTS system to provide multiple choices of unit for generating more accurate synthesized speech. **Approach:** This study proposes a phonetic context variable length unit selection MTTS system that is capable of providing more natural and accurate unit selection for synthesized speech. It implemented a phonetic context algorithm for unit selection for MTTS. The unit selection method (without phonetic context) may encounter the problem of selecting the speech unit from different sources and affect the quality of concatenation. This study proposes the design of speech corpus and unit selection method according to phonetic context so that it can select a string of continuous phoneme from same source instead of individual phoneme from different sources. This can further reduce the concatenation point and increase the quality of concatenation. The speech corpus was transcribed according to phonetic context to preserve the phonetic information. This method utilizes word base concatenation method. Firstly it will search through the speech corpus for the target word, if the target is found; it will be used for concatenation. If the word does not exist, then it will construct the words from phoneme sequence. **Results:** This system had been tested with 40 participants in Mean Opinion Score (MOS) listening test with the average rates for naturalness, pronunciation and intelligibility are 3.9, 4.1 and 3.9. **Conclusion/Recommendation:** Through this study, a very first version of Corpus-based MTTS has been designed; it has improved the naturalness, pronunciation and intelligibility of synthetic speech. But it still has some lacking that need to be perfected such as the prosody module to support the phrasing analysis and intonation of input text to match with the waveform modifier.

**Key words:** Text to speech, unit selection, concatenation, corpus-based speech synthesis, speech synthesis

## INTRODUCTION

There are three main characteristics that are inherent in a good quality text to speech synthesizer as shown in Fig. 1. The main factors that influence the quality are a good set of synthesis units, an efficient concatenation process that allows these units to be smoothly concatenated and finally the ability to synthesize natural prosodic content across the concatenated units in relation to the intended linguistic requirement[1].

Many popular speech synthesizers use concatenative methods to generate audible speech from text input[2]. Concatenative synthesis is a synthesis method that connects pre-recorded natural utterances to produce intelligible and natural sounding synthetic speech[3,4]. Concatenation can be actually accomplished by either overlap-adding stored waveforms or by reconstruction using method such as linear prediction or even formant synthesis. In concatenative systems, speech units can be either fixed-size diphones or variable length units such as syllables and phones[4].

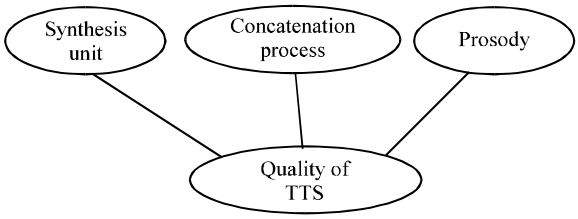


Fig. 1: Main Factors influence the quality of TTS[1]

**Corresponding Author:** Tian-Swee Tan, Faculty of Biomedical Engineering and Health Science, P11, Center for Biomedical Engineering, University Teknologi Malaysia, 81310 UTM Skudai, Johor DT, Malaysia Tel: +60127428412 Fax: +6075535430

Concatenative speech synthesis systems attempt to minimize audible discontinuities between two successive concatenated units[5]. In unit selection concatenative synthesis, a join cost is calculated that is intended to predict the extent of audible discontinuity introduced by the concatenation of two specific units[2]. One of the most important aspects in concatenative synthesis is to find correct unit length[6]. The selection is usually a trade-off between longer and shorter units[7]. With longer units, high naturalness, less concatenation points and good control of co-articulation are achieved, but the amount of required units and memory is increased. However, with shorter units, less memory is needed, but the sample collecting and labeling procedures become more difficult and complex. The units used for present concatenative synthesis systems are usually words, syllables, demi-syllables, phonemes, diphones, and sometimes even tri-phones.

Unit selection is the recent most used technique for concatenation and corpus based synthesis. It has improved the quality of synthetic speech by making it possible to concatenate speech from a large database to produce intelligible synthesis while preserving much of the naturalness of the original signal[2]. Figure 2 shows the concept of unit selection. It can be seen that the second unit is selected from a set of same unit by evaluating the distance between the adjacent unit.

Unit selection synthesis has the potential for higher quality and more natural sounding synthetic speech, but it also requires an algorithm to select at run time the most appropriate units available to construct the desired utterance[2]. The unit selection process ensures that the acoustic segments with matching left and right contexts are chosen.

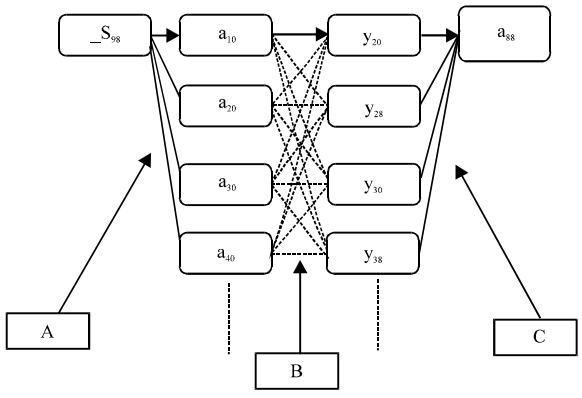


Fig. 2: Unit Selection

## MATERIALS AND METHODS

**Corpus-based TTS:** Corpus-based TTS system creates the output speech by selecting and concatenating units (e.g. Speech sounds or words) from a large (can be up to several hours long) speech database to select the units to be concatenated[8].

The idea of corpus-based or unit selection synthesis is that the corpus is searched for maximally long phonetic strings to match the sounds to be synthesized[9]. According to Nagy[10], as the length of the elements used in the synthesized speech increases, the number of concatenation points decreases, resulting in higher perceived quality. If the database offers sufficient prosodic and allophonic coverage, it is then even possible to generate natural sounding prosody without resort to signal manipulation.

This technique is capable to search for maximally long phonetic strings to match the sounds to be synthesized[9]. It uses a large inventory to select the units to be concatenated[4].

**Malay Waveform Generator Modules:** Malay Waveform Generator Modules, WGM is the new component that supports unit selection, concatenation smoothing and wave modifier as shown in Fig. 3. Concatenation smoothing utilizes the PSOLA's overlap-add function to smooth and remove the artifact click sound. Meanwhile wave modifier utilizes PSOLA to modify the duration, volume (loudness) and pitch level of the speech unit.

The WGM will first find the best match unit sequence, concatenate it, smooth the concatenation join and then modify the duration, pitch and volume.

**Malay Linguistic Transcription Module (MLT):** To access to the speech unit, a set of speech unit transcription file has to be designed. This transcription should descript the detail of the origin of the speech unit from its carrier sentence and the phonetic context.

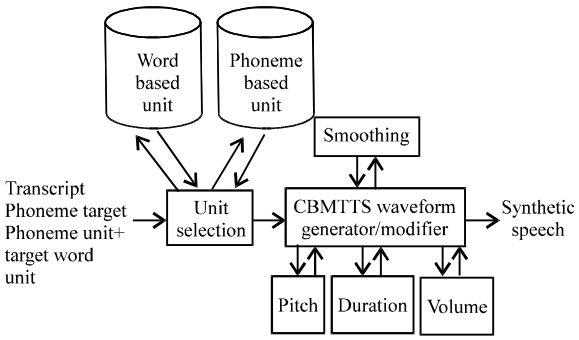


Fig. 3: The WGM Architecture of CBMTTS

| XX | C | P | N | x1 | x2 |
|---|---|---|---|---|---|

XX = Carrier sentence wavefile
C = Current phoneme
P = Previous phoneme
N = Next phoneme
x1 = Start index of wave from carrier sentence
x2 = End index of wave from carrier sentence

Fig. 4: Phonetic transcription for speech unit. (supposedly the target sequence transcription)

| XX | saya | P | N | x1 | x2 |
|---|---|---|---|---|---|

XX = Carrier sentence wavefile
Saya = Current phoneme
P = Previous phoneme
N = Next phoneme
x1 = Start index of wave from carrier sentence
x2 = End index of wave from carrier sentence

Fig. 5: Word Unit transcription. (supposedly the target sequence transcription)

Figure 4 shows the transcription format for speech unit. This transcription file descript the speech unit in term of its origin carrier sentence, phonetic context in term of previous and next phoneme, and wave location in original wave.

This module transcripts the input text into target unit sequence with phonetic context. As CBMTTS utilize word based concatenation and variable length unit selection, the MLT has been custom made to support two states transcription. The first state is to predict the sequence of target word unit with its linguistic context and the second state is to predict the sequence of smallest unit (in this case is phoneme) with its phonetic context.

This module transcripts the input text into target unit sequence with phonetic context. As MTTS utilizes word based concatenation and variable length unit selection, the MLT has been custom made to support two transcription states. The first state is to predict the sequence of target word unit with its linguistic context and the second state is to predict the sequence of smallest unit (in this case is phoneme) with its phonetic context.

Figure 4 shows the target word unit transcription format for word-selection and Fig. 5 shows the target phoneme unit transcription format for phoneme selection.

Table 1a: Comparison of concatenation point for single input word using different techniques

| Example word | saya | mempersembahkan |
|---|---|---|
| Phoneme sequence | s a y a | m e m p e r s e m b a h k a n |
| Phoneme based synthesis | 3 | 14 |
| Variable length unit selection | <3 | <14 |
| Word based synthesis | 0 | 0 |

(Remark: Assume the word exists in word based unit database)

Table 1b: Comparison of concatenation point for input sentences using different techniques

| Example word | Saya makan nasi. | Ali pergi ke sekolah dengan menaiki bas |
|---|---|---|
| Phoneme sequence | _s a y a _m a k a n _n a s i | _a l i _p e r g i _k e _ s e k o l a h _d e n g a n _m e n a i k i _b a s |
| Phoneme based synthesis | 12 | 30 |
| Variable length unit selection | <12 | <30 |
| Word based synthesis | 2 | 6 |

(Remark: Assume the all the words exist in word based unit database)

**Concatenation Point:** The formula of the phoneme based concatenation for single input word is as below:

$$T_{point} = T_{pho} - 1 \quad (1)$$

Where:
$T_{point}$ = Total concatenation point
$T_{pho}$ = Total phoneme

For example as in Table 1a, the word “saya” (which means “I”) has only 3 concatenation points and the word “mempersembahkan” (which means “presenting”) has only 14 concatenation points if using phoneme based synthesis. Meanwhile, it does not have any concatenation point if using word based concatenation (if the word is supported in the database). If the same word concatenated using variable length unit selection, it will require less concatenation point that phoneme based concatenation. This is because variable length unit selection may concatenate the word from a bigger unit than phoneme such as using two or more continuous phoneme.

The result will be the same for synthesizing the sentences. The total concatenation points for sentences input are shown in Table 1b. For example, the sentence “saya makan nasi” (which means “I eat rice”) has total 12 concatenation points if using phoneme based synthesis meanwhile has only 2 concatenation points if using word based synthesis (by assumption all words are supported in database). Same here, the variable length synthesis method will produce synthesized speech with less concatenation points than phoneme based synthesis method.

**Malay Word Based Concatenation System (MWBCS):** Since the creation of the speech corpus focuses on including the most frequent words, virtually all sentences requested for synthesis will contain portions that have no corresponding elements in the database. This means that the corpus must be constructed to include all possible phonemes in at least one version, but the more frequent ones in multiple contexts[10].

In a real application it may occur that words to be synthesized are not included in the database. In order to synthesize these missing words, we have chosen speech sounds to be the universal smallest units of the speech database.

Malay Word Concatenation Engine (MWCE) is a word construction unit selection engine that is custom made specifically for constructing non-exist word from phoneme unit in the database.

MWCE module as shown in Fig. 6 consists of phonetic context unit selection and spectral distance measure unit selection. Phonetic context unit selection will first match the transcript target unit with all existing speech unit in speech unit database. If more than 1 unit matches the phonetic context, then it will go to the second state of selection using the spectral distance measure.

**Corpus Based Speech Unit Concatenation Module (CBSUCM):** The speech unit concatenation module utilizes word-based concatenation. It receives input of target phoneme sequence and target word sequence, (with phonetic context) and concatenates the speech unit using word based concatenation engine. This is to reduce the concatenation point. Reduction of concatenation points through word based concatenation module will lead to reduction of artifact and increase of naturalness.

The unit selection process is shown in Fig. 6. Firstly, it will search through the word unit database by matching transcript target word sequence. If the word exists, then it will select the best match word unit. If the word does not exist, it will use the transcript target phoneme sequence to select from phoneme unit for forming word from. Once all word searching or forming is finished, then it will concatenate all the word units (either taken directly from database or construct from phoneme).

**Malay word based variable length unit selection:** Figure 7 shows the word-based variable length unit selection block diagram. Firstly, the phonetic context of target unit will be analyzed and match with speech unit in the database. If more than one speech unit match the phonetic context, then it will go through spectral distance measure to find out the best match unit with minimized distance.

**Concatenation engine:** As phoneme is the basic unit of speech, diphone, triphone or variable length of unit can be formed from it. Thus the engine has been modified to support variable length of unit instead of diphone.

It can select different length of unit from phoneme to diphone to triphone. Waveform concatenation module will first select the matching phoneme from speech database as shown in Fig. 7 then concatenate it to form new words. For example in Fig. 8, the word "suku" can be formed by the index 6 (_s), 7 (u), 19 (k) and 20 (u).

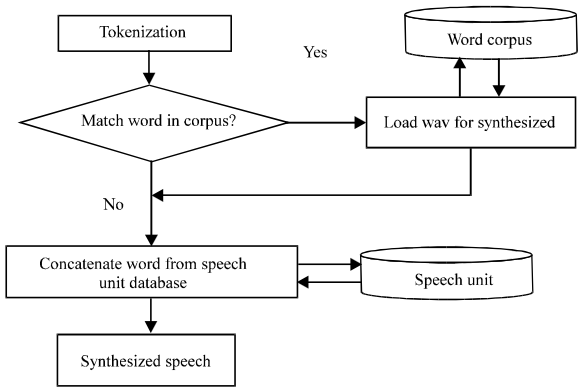


Fig. 6: Speech unit concatenation process

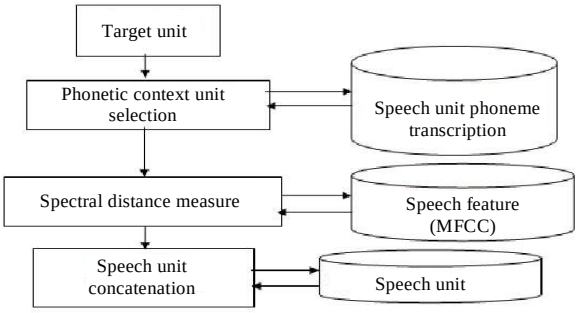


Fig. 7: Constructing non-exist word using malay word

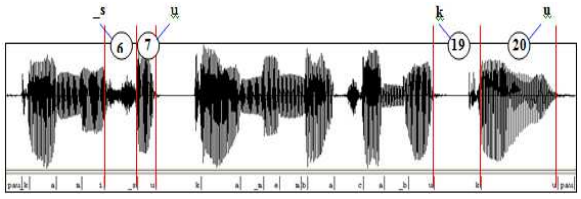


Fig. 8: Process of synthesizing new word from existing speech unit

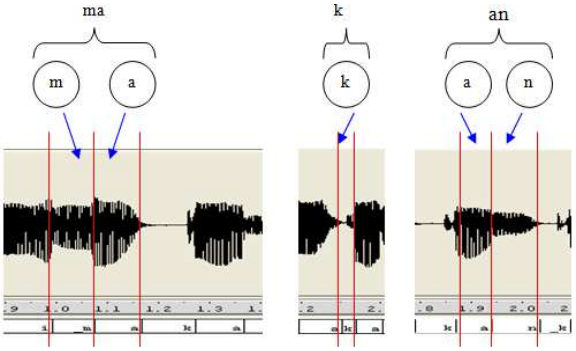


Fig. 9: Selection of speech unit for word concatenation

New word "suku' can be formed by concatenating diphones number 6, 7, 19 and 20.

Figure 9 shows another example of variable unit selection for the word "makan" formed from different sources such as carrier sentences 15, 103 and 1.

## RESULTS

The main interface of the system is shown in Fig. 10. This system can support for duration, volume and pitch modify that can be used for further study in intonation and expression speech synthesis in MTTS Synthesis system. Figure 11 shows another example of variable unit selection for the word "makan" form from few different sources such as carrier sentences 15, 103 and 1.

The MTTS has been tested via listening test. The listening test experiment is conducted via questionnaire be filled out by listener. The questionnaire was carefully designed to ensure the questions that provided are to evaluate the performance of CBMTTS as aforementioned.

The evaluation process endeavored to ascertain how the accuracy of the pronunciation of synthetic speech through Mean Opinion. MOS test has been used to test on certain categories of the performance of the system. The categories of the performance that have been tested on MTTS system are on naturalness, pronunciation, speed, stress, intelligibility, comprehensibility and pleasantness[11]. There are 3 categories of the output sound produced will be tested. These 3 categories are naturalness, pronunciation and intelligibility of the output sound. This listening experiment took 2 weeks to be completed and total of 40 listeners from different backgrounds took part in this experiment. Each participant took 30 min to complete the whole test including pretest. A set of computer PC Pentium IV 3GHz and headphone has been used for the test. Headphone is required for the experiment because the listener needs full concentration.


Fig. 10: The GUI for Malay Speech Synthesizer

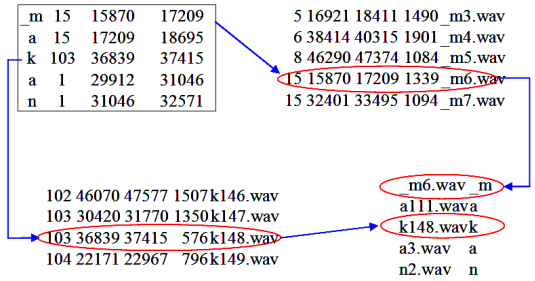


Fig. 11: Selection of speech unit for word concatenation

Table 2: Test words for MOS

| No | Word |
|---|---|
| 1 | *Tujuan* (purpose) |
| 2 | *Cakap* (say) |
| 3 | *Cukup* (enough) |
| 4 | *Ikan* (fish) |
| 5 | *Kertas* (paper) |
| 6 | *Kompas* (compass) |
| 7 | *Seterus* (next) |
| 8 | *Keselamatan* (safety) |
| 9 | *Kebanyakan* (many) |
| 10 | *Mempersembahkan* (perform) |

Table 2 shows the test words for MOS. It has been tested for 40 participants. From the MOS analysis, the average rate for naturalness is 3.9, average rate for pronunciation is 4.1 and average rate for intelligibility is 3.9. From the result obtained, average rates are in the upper middle of rate. The rate for naturalness, pronunciation and intelligibility are 4.4, 4.59 and 4.44 for the same word 3 (*cukup*). The lowest rate for naturalness, pronunciation and intelligibility are 3.37, 3.44 and 3.3 for the word 5 (*kertas*). The word 5 (*kertas*) get the lowest rate because of the bad pronunciation of "r" and "s".

Table 3 shows the comparison of performance between the Malay Speech Synthesis using diphone method and corpus-based. The diphone based concatenation method is the first version MTTS[12,13]. The new version of MTTS, which utilizes corpus based method has improved the quality of synthetic speech in naturalness, pronunciation and intelligibility.

Table 3: Comparison of performance between Malay diphone concatenation and Malay corpus based speech synthesis method

| Diphone | Corpus based | Concatenation |
|---|---|---|
| Naturalness | 3.11 | 3.9 |
| Pronunciation | 3.49 | 4.1 |
| Intelligibility | 3.50 | 3.9 |

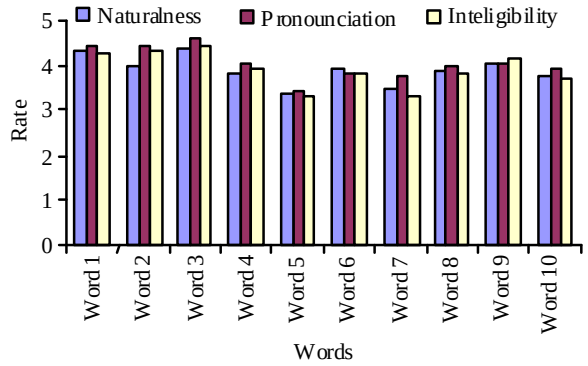


Fig. 12: Mean opinion score result

## DISCUSSION

Through this project, a very first version of Corpus-based MTTS has been designed. It has complete platforms of text preprocessing (tokenizer, normalize), linguistic analysis (word tagging and phonetizer) and waveform generator (unit selection, concatenation, smoothing and waveform modifier). The system has been verified through listening test and applied in speech therapy and sign language to speech system, though it still has some lacking that needs to be perfected such as the prosody module to support the phrasing analysis and intonation of input text to match with the waveform modifier. Besides that, the predication of pronunciation also needs to be perfected with more rules support or pronunciation dictionary. This will require the collaboration with linguistic expert to create the rule and design the system that can support for automatic prosody and intonation generation. But since the waveform modifier has been designed to support the pitch, duration and volume control, it would not be difficult to create the automatic intonation prediction module that can generate the synthetic speech with intonation support in future.

## CONCLUSION

This research has proposed a phonetic context variable length unit selection method for Corpus-based MTTS. It also utilizes word-based concatenation which provides higher accuracy of word selection through word corpus and phoneme corpus. Through the database transcription design it can preserve the phonetic context of the speech units either in word corpus or phoneme corpus. Thus, it provides the flexibility for choosing a string of continuous phoneme through its preserved phonetic context. It has been proven to improve the quality of MTTS synthesis in its pronunciation, intelligibility and naturalness.

## ACKNOWLEDGEMENT

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

## REFERENCES

1. Low, P.H. and S. Vageshi, 2002. Synthesis of unseen context and spectral and pitch contour smoothing in concatenated text to speech synthesis. Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 13-17, IEEE Xplore Press, Orlando, Florida, pp: I-469, I-472. DOI: 10.1109/ICASSP.2002.1005778.
2. Ann, K.S. and D.C. Alistair, 2004. Data-driven perceptually-based join costs. Proceeding of the 5th ISCA ITRW on Speech Synthesis, June 14-16, ISCA, Carnegie Mellon University, Pittsburgh, pp: 49-54. http://citeseer.ist.psu.edu/634094.html.
3. Andersen, O., N.J. Dyhr, I.S. Engberg and C. Nielsen, 1998. Synthesizing short vowels from their long counterparts in a concatenative based text-to-speech system. Proceeding of the 3rd ESCA Workshop on Speech Synthesis, November 26-29, ESCA, Australia, pp: 147-151. http://www. iscaspeech. org/archive/ssw3/ssw3_165.html.
4. Hasim, S., G. Tunga and S. Yasar, 2006. A corpus-based concatenative speech synthesis system for Turkish. Turk. J. Elect. Eng. Comput. Sci., 14: 209-223. http://www.cmpe. boun.edu.tr/~gungort/papers/A%20Corpus-Based %20Concatenative%20Speech%20Synthesis%20S ystem%20for%20Turkish.pdf.
5. Stylianou, Y. and A.K. Syrdal, 2001. Perceptual and objective detection of discontinuities in concatenative speech synthesis. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 7-11, IEEE Xplore Press, USA., pp: 837-840. DOI: 10.1109/ICASSP.2001.941045.

6. Lewis, E. and T. Mark, 1999. Word and syllable concatenation in text-to-speech synthesis. In: 6th European Conference on Speech Communications and Technology, September, ESCA, Australia, pp: 615-618. Doi: http://citeseer.ist.psu.edu/eric99word.html
7. Black, A. and N. Campbell, 1995. Optimising selection of units from speech databases for concatenative synthesis. Proceeding of Eurospeech, September 18-21, Eurospeech, Madrid, Spain, pp: 581-584. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.40.4014.
8. Hasim, S., G. Tunga and S. Yasar, 2006. A corpus-based concatenative speech synthesis system for Turkish. Turk. J. Elect. Eng. Comput. Sci., 14: 209-223. www.cmpe.boun.edu.tr/~hasim/hasimCV.pdf.
9. Joakim, N., K. Heiki-Jaan, M. Kadri and K. Mare, 1996. Designing a speech corpus for estonian unit selection synthesis. Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007, May 24-26, Tartu, pp: 367-371. http://dspace.utlib.ee/dspace/handle/10062/2605.
10. Nagy, A., P. Pesti, G. Németh and T. Bőhm, 2005. Design issues of a corpus-based speech synthesizer. Hungarian J. Commun., 6: 18-24. www.cc.gatech.edu/~pesti/pubs/ht_cikk_en_2005.pdf.
11. Hirst, D., A. Rilliard and V. Aubergé, 1998. Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis. Proceeding of the 3rd ESCA/COCOSDA Workshop on SPEECH SYNTHESIS. November 26-29, ISCA Press, Jenolan Caves, Blue Mountians, NSW Australia, pp: 1-4. http://www.isca-speech.org/archive/ssw3/ssw3_001.html.
12. Tan, T.S., S. Hussain and A. Hussain, 2003. Building malay diphone database for malay text to speech synthesis system using festival speech synthesis system. Proceeding of the International Conference on Robotics, Vision, Information and Signal Processing, January 22-24, ROVISP, pp: 634-648. http://www.eng.ukm.my/jkees/Prosiding_2003.htm.
13. Tan, T. S., 2004. The Design and Verification of Malay Text to Speech. M. Eng. Thesis, University of Technology Malaysia, Skudai, Malaysia. http://sps.utm.my/sps/Academic%20Resources/abstract-of-thesis/2004/tan-tian-swee.