# NATURAL SOUNDING STANDARD MALAY SPEECH SYNTHESIS BASED ON UTMK EBMT ARCHITECTURE SYSTEM

**SABRINA TIUN**

**UNIVERSITI SAINS MALAYSIA**

**2011**

# NATURAL SOUNDING STANDARD MALAY SPEECH SYNTHESIS BASED ON UTMK EBMT ARCHITECTURE SYSTEM

**by**

**SABRINA TIUN**

**Thesis submitted in fulfilment of the requirements**
**for the degree of**
**Doctor of Philosophy**

**January 2011**

# ACKNOWLEDGEMENTS

Alhamdulilah. There so many people I would like to express my greatest gratitude for the completion of this thesis. My respected supervisors, Prof. Dr. Rosni Abdullah and Prof. Madya Dr. Tang Enya Kong, thank you so much for the faith, support, guide, advice, motivation and encouragement. Not only both of you are my role-models, but also the best "boss" anyone could ask for. I would also like to convey my deepest appreciation to the examiners of this thesis for their comments and suggestions, namely Prof. Dr Tengku Mohd Tengku Sembok, Prof. Dr Ahmad Zaki bin Abu Bakar and Dr. Tan Tien Peng.

I am grateful to the Unit Terjemahan Melalui Komputer, School of Computer Sciences and Instutite of Postgraduate Studies (IPS), for any help and facilities given. Deepest gratitude to USM for granted the tuition fee exemption throughout my PhD course.

I owe a huge thanks to many people who have been supportive and helpful, especially to Norliza Hani and Anuar for their help in preparing the speech data, Siti Khaotijah for enlightening me on the world of linguistics, to Hong Hoe for the his guidance on EBMT architecture and to Lian Tze for sharing the latex template of USM Thesis. To all the people who have involved in the test evaluation, thank you so much for spending your precious time on the test.

I would also like to express sincere thanks to my colleagues and my fellow graduate friends for their support, friendship and encouragement during my PhD journey. I am blessed to be surrounded by wonderful people like you guys.

To Amil, Adi, Aidan and Anuar, thank you for the love and understanding. Especially to dear Anuar who have played so many roles in the completion of this thesis, I owe you so much. And not forgetting, thank you so much to both of my parents and my siblings for your love and encouragement, and especially to my father who always believe in me.

All Praises to Allah the Almighty, for granting my dream becomes reality.

Sabrina,
January,2011.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

**IPS** Institut Pengajian Siswazah

**PPSK** Pusat Pengajian Sains Komputer

**USM** Universiti Sains Malaysia

**UTMK** Unit Terjemahan Melalui Komputer

**EBMT** Example-based Machine Translation

**MT** Machine Translation

**ANOVA** ANalysis Of VAriance

**MOS** Mean Opinion Scores

**TTS** Text-to-Speech

**IVR** Interactive Voice Response

**MOS** Mean Opinion Scores

**SSTC** Structured String-Tree Correspondence

**S-SSTC** Synchronous Structured String-Tree Correspondence

**UTMK-MSS** Unit Terjemahan Melalui Komputer Malay Speech Synthesiser

**SM** Standard Malay

**LP** Linear Prediction

**TD-LP** Time-Domain Linear Prediction

**ATR** Advanced Telecommunication Research Institute International

**LSFs** Linear Spectrum Frequencies

**MFCC** Mel-frequency cepstrum coefficients

**AT&T** American Telephone & Telegraph

**HNM** Harmonic plus Noise Model

**PSOLA** Pitch Synchronous Overlap Add

**HMM** Hidden Markov Model

**MSS** Malay Speech Synthesiser

**MIMOS** Malaysian Institute of Microelectronic Systems

**UTM** Universiti Teknologi Malaysia

**POS** Part-Of-Speech

**NL** Natural Language

**ITU** International Telecommunication Union

# SINTESIS PERTUTURAN BAHASA MELAYU STANDARD BERBUNYI SEMULAJADI BERDASARKAN SENIBINA SISTEM UTMK EBMT

## ABSTRAK

Matlamat utama kajian tesis ini ialah untuk membina sintesis pertuturan bahasa Melayu yang berbunyi semulajadi. Matlamat ini dipilih berdasarkan jenis aplikasi sintesis pertuturan yang diperlukan oleh pasaran industri, iaitu aplikasi domain-terhad. Aplikasi sintesis pertuturan domain-terhad ialah aplikasi yang mempunyai jumlah kosa kata yang terhad (kurang fleksibel) tetapi memerlukan suara pertuturan berbunyi semulajadi berkualiti tinggi. Berdasarkan evolusi teknik sintesis pertuturan, penggunaan unit pertuturan yang semulajadi tanpa mengaplikasi sebarang pemprosesan isyarat merupakan satu teknik terkini yang mampu mengeluarkan pertuturan sintetik yang tinggi kualiti bunyi semulajadinya. Oleh kerana itu, kami memilih untuk menggunakan teknik sintesis pertuturan yang mana titik pencantuman unit pertuturan dan manipulasi prosodi dapat dielakkan atau dikurangkan. Teknik ini diimplementasikan dengan menggunakan saiz unit pertuturan yang besar dan memperbanyakkan contoh pada tiap-tiap jenis unit pertuturan. Namun, apa yang menjadi persoalan ialah bagaimana untuk memilih contoh unit pertuturan yang sesuai untuk ayat yang hendak disintesiskan? Dalam kajian tesis ini, kami menyelesaikan permasalahan dalam pemilihan unit pertuturan menggunakan pendekatan sintesis pertuturan berdasarkan-korpus, yang mana kami menggunakan penghurai berdasarkan-contoh dari sistem

Example-based Machine Translation (EBMT) Unit Terjemahan Melalui Komputer (UTMK), dan menjadikan struktur pokok sintaksis-prosodi sebagai perwakilan bagi korpus pertuturan. Terdapat tiga kerja penyelidikan yang penting dilakukan dalam tesis ini, iaitu membina korpus pertuturan sintaksis-prosodi, mengadaptasi senibina sistem EBMT UTMK untuk menghasilkan pesintesis pertuturan dan mencadangkan pencantuman unit subperkataan yang tiada bunyi herotan. Semua kerja penyelidikan ini adalah untuk membina model alat sintesis pertuturan bahasa Melayu yang berbunyi semulajadi dan mempunyai sedikit sifat fleksibel. Kami menilai prestasi pendekatan sintesis pertuturan kami dengan melakukan ujian MOS diubahsuai, analisa prosodik-akustik dan ujian kelancaran. Berdasarkan analisa ujian statistik ANOVA dan T-test, output dari sintesis pertuturan kami secara signifikannya adalah lebih berbunyi semulajadi berbanding dengan output sintesis dari sistem sintesis pertuturan bahasa Melayu standard yang lain. Bagaimanapun, ia masih kurang bunyi semulajadinya jika dibandingkan dengan pertuturan yang semulajadi.

# NATURAL SOUNDING STANDARD MALAY SPEECH SYNTHESIS BASED ON UTMK EBMT ARCHITECTURE SYSTEM

## ABSTRACT

In this research work, we make natural sounding speech synthesis as the main goal. This goal was chosen following the type of demanded speech synthesis application systems by the industrial market; the limited-domain speech synthesis application systems. The limited-domain speech synthesis application system has restricted number of vocabularies (less flexible) but requires a highly natural sounding of speech synthesis. Based on the evolution of speech synthesis technique, one can conclude that using natural speech units without applying any signal processing is the technique to produce the most natural sounding of synthetic speech. As such, we opt to use a synthesis technique that avoids (or lessen) the concatenation points and prosodic manipulation process. The technique is implemented by using larger chunk of synthesis unit and making as much as possible the instances of one particular type of speech unit. However, the big question is how to choose the right instances of speech units for the targeted sentence? In this thesis, we address the speech unit selection problem by using corpus-based speech synthesis approach, in which, we use the example-based parser of Unit Terjemahan Melalui Komputer (UTMK) Example-based Machine Translation (EBMT) system, and speech corpus represented by a syntax-prosody tree structure. There are three significant research works conducted in this

thesis; viz. the creation of syntax-prosody speech corpus, adapting the UTMK EBMT system architecture to create speech synthesiser system and proposing the non-audible distortion of subword unit concatenation. These contributions serve for one goal, which is to build a natural sounding Malay speech synthesiser model with a little bit of flexibility characteristic. We assess the performance of our speech synthesis approach by conducting a modified MOS test, prosodic-acoustic analysis and smoothness test. Based on the statistical analysis using ANOVA and T-tests, significantly our synthesis output was perceived more natural than the other synthesis output of Standard Malay speech synthesiser systems, but, less natural than the natural speech.

# CHAPTER 1

# INTRODUCTION

## 1.1 Research Motivation

Human communication with machines has tremendously improved over the years and this is greatly due to the huge improvements on machine capability of doing the listening and talking with human. Such interfaces are built using the speech processing technology. For listening, a machine will use speech recognition processing to do the task and for talking, the machine will use speech synthesis processing. The component of a machine that carries out the speech synthesis processing is called a speech synthesiser.

Speech synthesiser is required in almost any automated information or services systems and multimedia application systems. Even though speech synthesiser is highly demanded in those application systems, but, only a few companies are using speech synthesisers in or as their systems interfaces. This is mainly due to the speech synthesis voice quality that has not yet met up to the end-users' expectations. Therefore, improving the voice quality or also known as producing the natural sounding synthetic speech is one of the most appealing research areas in speech synthesis processing.

In a simple sentence, speech synthesis can be described as the process of producing artificial human speech or also known as synthetic speech. The ultimate goal of inventing speech synthesiser system is to design and program a machine to talk like

human. In order for the synthetic speech not to be misunderstood by human either on the meaning interpretation or pronunciation, the speech synthesiser must produce a high quality synthetic utterance. A high quality speech synthesiser should have these two qualities; intelligibility and naturalness (Yi and Glass, 1998; Klabbers, 2000).

In speech synthesis, intelligibility is always referred to as how well-rendered the speech output is produced by the speech synthesiser to be clearly perceived commonly by human hearing (Tatham and Morton, 2005). Naturalness, on the other hand, is always perceived as the ability of a speech synthesiser to produce synthetic speech similar to the human speech. Yet, the definition on naturalness is still much debated. Hawkins et al. (2000) stated that naturalness should be defined as how easy to understand the synthetic as we understand human speech. The same point of view is shared by Hess (2008) which argues that naturalness does not mean that synthetic speech should sound like real human speech but rather as a voice that is easy to listen to, given in the same environment as natural speech.

An ideal speech synthesiser should be flexible without jeopardising the qualities of intelligibility and naturalness (Klabbers, 2000). The term flexibility here refers to the capability of a speech synthesiser to synthesise any arbitrary sequence of input word or sentence.

However, Yi and Glass (1998) and Klabbers (2000) claimed that flexibility and naturalness is actually the trade-off of building speech synthesiser. In order for a synthesiser to favour flexibility, somehow it has to sacrifice its naturalness and also vice-versa. The current speech synthesiser systems are built based on prioritising flexibility before naturalness (Klabbers, 2000; Cox et al., 2000). Alternatively, a

synthesiser can be built by maintaining naturalness and then later proceed to flexibility (Yi and Glass, 1998).

The reason why we choose natural sounding as the first priority in building a speech synthesiser is because most of the application systems in the industrial companies require the voice of the speech synthesiser to sound naturally. A natural sounding voice is crucially needed since the voice is aimed to replace human's natural voice. In the other hand, the flexibility of the speech synthesiser is not as crucial as its naturalness quality since most of these application systems are limited-domain that have restricted number of vocabulary.

Looking at the Nusuara[1]'s website at `http://www.nusuara.com/solutions/`, most of Nusuara speech interface application systems are call centres, automated telephony services and interactive voice response (IVR). Based on that, we can conclude the trend of Malaysian market is also on limited-domain speech synthesis applications. As we mentioned before, the limited-domain speech synthesis application systems have restricted number of vocabularies, which is less flexible, but require a highly natural sounding of speech synthesis. Therefore, we believe that our research work is on the right track in fulfilling the demand of the Malaysian industrial companies that intend to use speech technology as part of their business solutions.

Thus, the main objective of this thesis is to build a natural sounding Standard Malay (SM) speech synthesiser model with a little bit of flexibility characteristic.

[1] A Malaysian company which specialised in developing application based on speech technologies.

## 1.2 Research Problem and Solution Approach

As suggested by Stöber et al. (1999), the fast and cheap way to provide a highly natural sounding voice output for a speech synthesis application is by having none or less concatenation points and prosodic manipulation process. The approach taken by Stöber et al. (1999) was based on the idea of using word as primary synthesis unit and making as much as possible the instances for one particular word unit. Although, this could mean an indefinite number of speech units, but, most of the speech synthesis applications are limited domain application systems, therefore the size of the vocabulary is small. The main question now is how to choose the right word unit (sound) for the target sentence? Besides, domain speech application system should also have the mechanisms to handle novel word yet maintains the naturalness of its speech output.

In this thesis, we address the word selection problem by using a corpus-based speech synthesis approach. However, instead of using the unit selection approach, we adapt the example-based approach of Unit Terjemahan Melalui Komputer (UTMK) Example-based Machine Translation (EBMT) system (Ye, 2006) and our speech corpus is represented by syntax-prosody tree structures.

In the syntax-prosody speech corpus, every sentence is represented with a single dependency syntax-prosody tree. Each of the nodes in the tree will be annotated with prosodic information and aligned with sound (word synthesis unit).

Therefore, given an input sentence, the example-based parsing will parse the sentence into a dependency syntax-prosody tree. In order to generate the synthetic

utterance of the input sentence, every speech unit aligned with the nodes of the parsed tree will be retrieved and concatenated.

We choose to adapt the UTMK EBMT processing technique to select speech unit and represent our speech corpus in syntax-prosody tree structure because of the following reasons:

- The UTMK EBMT system is a deliverable product that use example corpus as its knowledge-base. Theoretically, a corpus-based approach is possible to be applied on any target media as long as the corpus data is large and correctly prepared. Here, even though UTMK EBMT is using example-based approach for translating text, but, we believe that by using its architecture and its example-based parser, we can build a corpus-based speech synthesiser.
- By using syntax-prosody tree representation, speech units are expected to be selected more accurately compared to the standard algorithm of corpus-based speech synthesis, the unit selection. The main idea behind this is appropriate speech units are likely to be retrieved from appropriate context (Möbius, 2000). Retrieving the speech units using the example-based parser of UTMK EBMT, which is by constructing a tree from subtrees of target text, we implicitly select the most appropriate speech unit. The most appropriate speech units here defined as the speech units that are co-occurring together, or in other words, speech units that come from the same wave file.

The use of the example-based parser and the syntax-prosody speech corpus is to avoid prosodic mismatch of the concatenated selected speech units. In order to ensure

segmental mismatch does not happen, we choose speech unit that does not have strong co-articulation on its neighbouring phonemes. Therefore we only choose three types of speech units; phrase, word and subword.

Phrase unit defined in this thesis is a segment in a sentence in which, if it is a spoken phrase, it is distinguished with a brief pause. In a written form, the phrases in a sentence are delimited by comma (,), semicolons (: and ;). As for word unit, it is a segment or strings delimited by blank spaces in a sentence. Thus, the written form of word 'layang-layang (kite)' is one word unit (instead of just 'layang'), even though in a spoken form a very brief pause may occur between the string 'layang'. Subword unit, in the other hand, is a syllable or syllables extracted from the word unit.

The subword unit is the synthesis unit we use to generate novel word. However, the subword unit concatenation can potentially degrades the naturalness of the generated speech. Therefore, we avoid the segmental mismatch of subwords by only synthesising novel word from the list of subwords units that do not cause audible distortion. The subword list is created based on the result of phoneme adjacency analysis.

Therefore, by avoiding or minimising both prosodic and segmental mismatch, we expect our generated speech output will sound more natural than the speech output generated by the existing Standard Malay (SM) text-to-Speech (TTS) systems.

## 1.3 Research Objective, Contributions and Limitations

The objective of this thesis is to choose the right speech units of a target text that eventually the concatenation of those speech units will make a natural sounding

synthetic voice. Since we also want to introduce a little flexibility without jeopardising naturalness quality, we try to find the speech unit smaller than word unit that sounded natural without applying any signal processing technique.

The objective is aimed to be achieved by adapting the UTMK EBMT architecture. By reusing and remodifying the UTMK EBMT modules, a natural sounding Malay speech synthesiser is built. Several works were carried out in order to do the adaptation and the description on each of the work is explained below:

- Using the existing resources of UTMK EBMT system; the Synchronized-String Structured Correspondences (S-SSTC) annotation and the syntactic tree of the examples in EBMT knowledge-base, we build a syntax-prosody speech corpus. In the work, we propose a single tree representation of syntax and prosodic features to represent speech and the kind of suitable prosodic information to be integrated with the syntactic tree structure.
- The UTMK EBMT system concerns with the accuracy to choose the right corresponding translated word or words, whereas the objective of our work is to choose the right instance of target speech unit. By adding new features, re-ranking features priority and removing irrelevant features in the example-based parser module, adding a synthesiser module and changing the bilingual dictionary module into a subword lookup module, we propose a new model of corpus-based speech synthesis system.
- We propose an idea to use subword speech units to generate novel sound. The subword synthesis units were carefully created based on a Standard Malay (SM) phoneme adjacency analysis. As far as we concern, the type of subword unit

that we propose has been never implemented in any SM speech synthesis system since phoneme adjacency analysis of SM has never been implemented before.

In this thesis, our focus concerns on naturalness and any problem regards to other aspects of speech synthesis performance viz. full flexibility, efficiency and robustness, will be left out for future work.

## 1.4 Thesis Outline

This thesis is organised in six chapters. This current chapter provides the introduction starting with the motivation on why we choose the "alternative" approach in modeling speech synthesis system. Afterwards, we itemise what are the contributions of this thesis, after we briefly discuss on the research problem and proposed solution of this thesis. The rest of this thesis is written out as follows:

*Chapter Two* is the chapter which describes the background of our research problem, the unnaturalness in speech synthesis. In the chapter, the root problems that cause the unnaturalness in previous and current speech synthesis were highlighted, together with the existing SM speech synthesis systems.

*Chapter Three* explains all theories and concepts related to the proposed solution of our research problem. Implicitly, it is divided into two major parts; First, it concerns on the literature review on syntax-prosody theoretical relationship and how the theories have been implemented in speech synthesis processing. Second, it is about the description on the Structured String-Tree Correspondence (SSTC) formalism and the current version of UTMK EBMT architecture.

*Chapter Four* is the implementation chapter. The chapter contains the description on how we build our syntax-prosody speech corpus and the subword unit lookup. We also include our little study on SM phoneme adjacency analysis as the prerequisite to build the subword unit lookup. At the end of this chapter, we describe the process of our Malay speech synthesiser, named as UTMK-Malay Speech Synthesiser (UTMK-MSS), based on the built speech corpus and subword unit lookup.

*Chapter Five* is about the evaluation tests on the speech quality of UTMK-MSS output. For speech synthesis mainly based on word units and did not involve subword, we used the modified MOS subjective evaluation and prosodic-acoustic analysis test. For subword unit synthesis, we use the smoothness test, as the test for subjective evaluation on the joint of the subwords.

Finally, in *Chapter Six* we summarise and conclude our research work. Not forgetting, we also lay out the future expansion of this research work. We are ambitious to see this speech synthesiser to be more natural by adding semantic knowledge, to make this synthesiser more flexible and to extend this research work into other research applications, e.g speech-to-speech machine translation.

# CHAPTER 2

# BACKGROUND

## 2.1 Introduction

In this chapter, we will elaborate on the factors that degrade the naturalness of synthesised utterance produced by concatenative speech synthesis approach and describe in detail on how concatenative speech synthesis approaches tackle the naturalness degradation problem. We also give a brief overview of the techniques used by the existing SM speech synthesisers, and end this chapter with a summary.

## 2.2 Degradation of Naturalness Quality in Concatenative Speech Synthesis

A Text-to-Speech (TTS) system is a complete system that generates waveform output from an input text. Fig. 2.1 shows the basic blocks of a TTS system. It mainly consists of four blocks: text analyser, phonetic analyser, prosodic analyser and speech synthesiser. Starting with the text input, the text analyser will do text normalisation and document structure on the text input. Text normalisation performs task like converting numerical symbol into alphabets; for example, the orthographic '1' converted into 'one'. Document structure detection, on the other hand, is responsible for chunking out text into sentences and words.

The normalised text will then be converted from orthographic symbols into

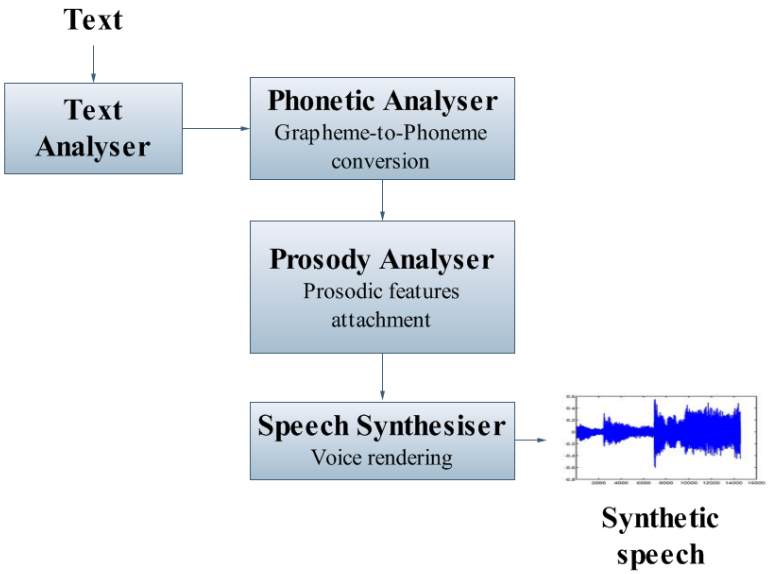


Figure 2.1: Overview of TTS general system (adapted from Huang et al. (2001)).

phonetics symbols by the phonetic analyser module, and this can be done by using a rule-based or dictionary lookup. In order to make the phonemics text sounded with the right intonation, the prosodic analyser will associate the phonetic symbols with correct prosodic values; pitch, duration and loudness. Finally, the speech synthesiser module will decode the phonetic symbols (that associates with prosodic features) into a speech output.

In general, the speech synthesiser or synthesis module can be categorised into three types: Articulatory synthesis, formant synthesis and concatenative synthesis (Huang et al., 2001). Articulatory and formant synthesis models are synthesis-by-rule model (Dutoit, 1997; Huang et al., 2001). Both of these approaches attempt to some degree to model how human produces speech.

Concatenative synthesis, on the other hand, is a data-driven synthesis model (Huang et al., 2001). It stores pieces of real speech chunks and concatenates the speech chunks to produce utterance. Therefore, obviously both articulatory and formant synthesis models are far more complex and difficult to build compared to concatenative synthesis model, due to the vast knowledge that need to be handled.

Despite the simplicity, concatenative speech synthesis is the current trend of modeling speech synthesis because of its capability of producing high quality of naturalness. The reason is because the speech chunks or unit are extracted from natural speech and therefore, synthesis of these speech units potentially produces a natural sounding speech output.

Unnaturalness in concatenative speech synthesis can be caused mainly during the preparation of the speech corpus and also during the process of synthesis. The processes of speech corpus preparation that can cause unnaturalness are mainly during the recording and the segmentation and labelling process.

Since concatenative speech synthesis takes synthetic utterance from natural speech, eventually the quality of the synthetic utterance also depends on the quality of the recorded speech. The voice for recording (professional speaker or not) and condition of the recording room need to be considered before the recording takes place.

Segmentation of the speech corpus into the chosen synthesis unit is also considered as the important process to ensure high quality of synthetic utterance. Since automatic segmentation is necessary in preparing large corpus, some errors do occur and this can degrade the synthesised utterance. According to Kominek et al. (2003), some of these

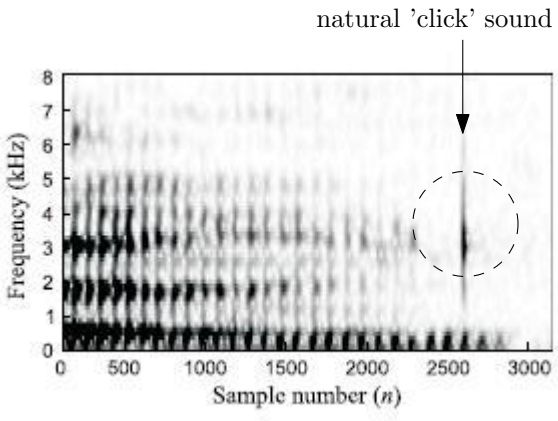


Figure 2.2: A spectrogram of noise sound 'click'.

errors are labels misalignment and segmentation, noise treated as speech sound and transcriptions errors. Fig. 2.2 shows the example of a noise sound 'click' in a recorded speech that is mistaken as a speech sound.

Since speech corpus is prepared offline, thoroughly checking and manually editing can ensure the speech corpus condition will be in excellent quality. Therefore, the most crucial part is to overcome the naturalness degradation during the synthesising process.

According to Klabbers and Veldhuis (2001), a high naturalness in concatenative speech synthesis can be achieved if both segmental and prosodic qualities are high. Prosodic quality is high when there is no mismatch of the prosodic features. These prosodic features can be in terms of pitch, duration and loudness. Among the three features, pitch is the most sensitive. A slight mismatch of pitch can obviously degrade naturalness (Dutoit, 2008) and that is why prosodic mismatched sometimes simplified as pitch mismatch (Huang et al., 2001). On the other hand, the segmental quality is high if a joint speech units does not introduce audible discontinuities. This audible discontinuity is largely contributed by the spectral mismatch (Klabbers and Veldhuis, 2001).

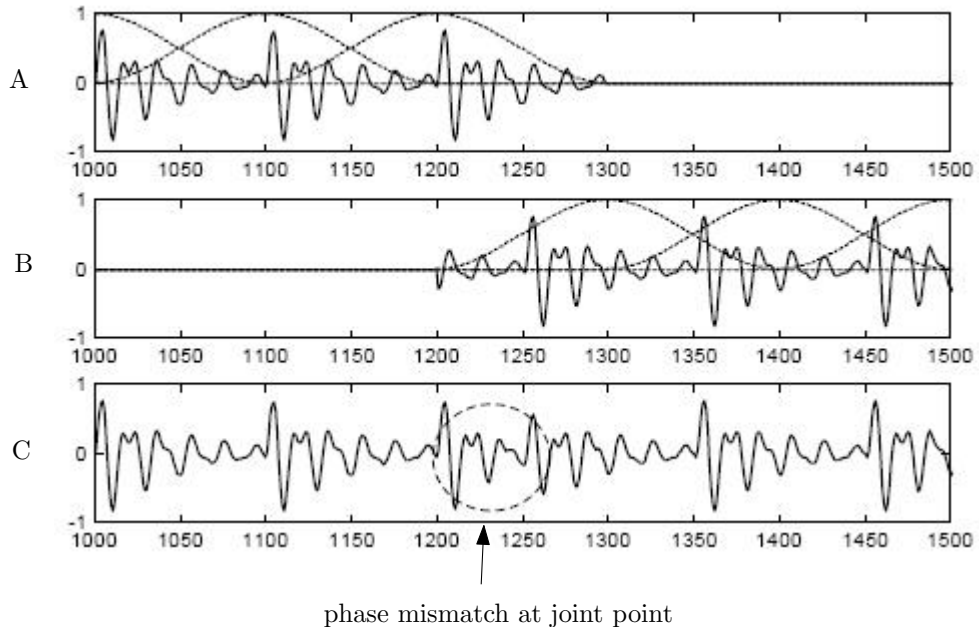


Figure 2.3: Waveform C contains a phase mismatch at the joint point of two sounds, A and B (adapted from Huang et al. (2001)).

According to Dutoit (2008), at the concatenation points, three mismatches of acoustic features can occur, and they are phase mismatch, pitch mismatch and spectral mismatch.

**Phase mismatch**. A phase mismatch occurs when the waveform of two synthesis unit are similar but the overlapped added frames are not centered (Dutoit, 2008). According to Stylianou (1998), this mismatch perceived as "garbled" sound in degraded speech quality and only takes effect at the concatenations point of voiced synthesis units. For example, in Fig. 2.3, the waveform A and B are similar yet at the mismatched region of waveform C, the length and shape of the circled waveform are not similar compared to both of its left and right regions. The mismatched region is the example of phase mismatch.

**Pitch mismatch**. A pitch mismatch occurs when two synthesis unit pitch values are different at the concatenation points (Dutoit, 2008). Since it can cause

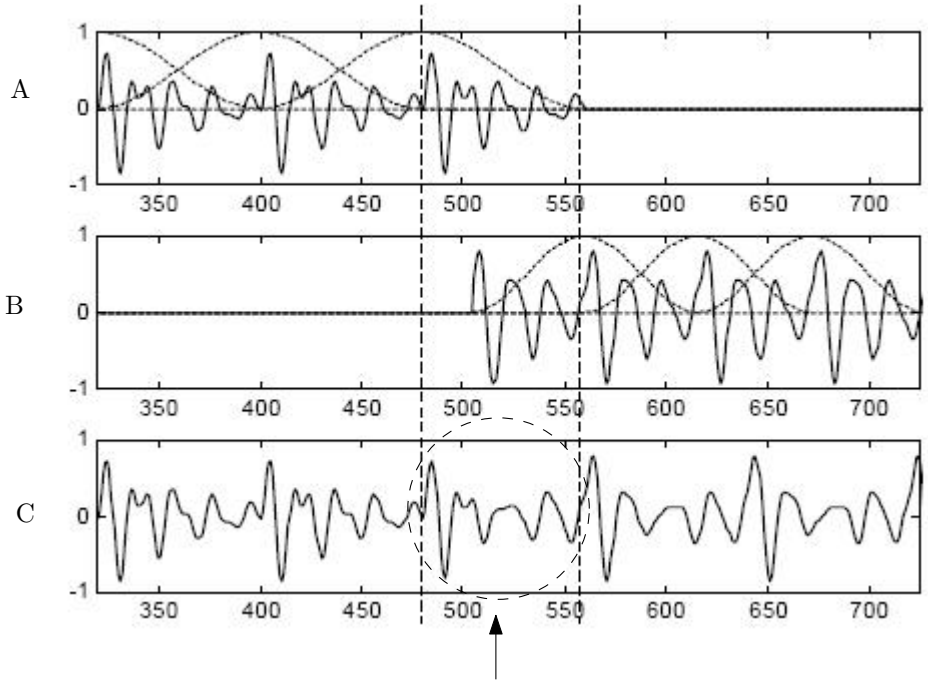


Figure 2.4: Waveform C contains a pitch mismatch at the joint point of sound A and B (adapted from Huang et al. (2001)).

audible discontinuity, therefore, it can degrade the naturalness quality of synthesised utterance.In Fig. 2.4, at the circled region of waveform C, audibility discontinuity can be perceived because of pitch mismatch. The pitch mismatch occurs because the waveform of A and B are not similar, even though the concatenation of the two waveforms is smooth.

**Spectral mismatch**: A spectral mismatch is a mismatch of spectral tilt, formant frequencies and bandwidths values at the concatenation points of two joint speech units (Stylianou, 1998). The mismatch can cause audible discontinuities and leads to the degradation of naturalness. A spectrogram in Fig.2.5 shows a spectral mismatch caused by the different values of formant frequencies at F2, which the left pointed F2 is higher than the right F2, across the boundary of two phonemes (each phoneme are taken from different contexts).

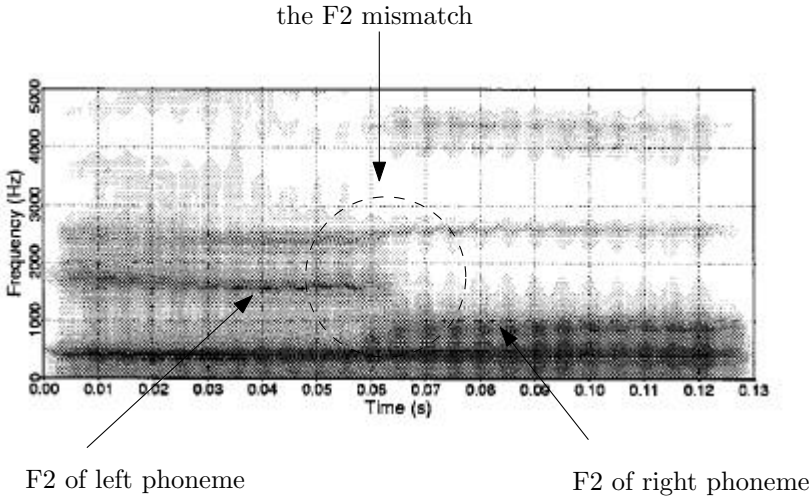


Figure 2.5: A spectrogram picture shows a spectral mismatch (adapted from Klabbers and Veldhuis (2001)).

## 2.3 Concatenative Speech Synthesis and Naturalness Quality

In this section, we look at the approaches in speech synthesis; particularly on concatenative speech synthesis. Since our concern is on the naturalness aspect of these approaches, therefore, we will meticulously lay down the problems of previous speech synthesis on naturalness and how the current concatenative speech synthesis is designed to overcome these problems.

The segmental (e.g. spectral) and prosodic (e.g. pitch) mismatch problem are handled quite differently based on which model the concatenative speech synthesis is built upon. Concatenative speech synthesis techniques can be classified into: Synthesis with fixed inventory, corpus-based synthesis and HMM-based synthesis (Dutoit, 2008). Thus, we will briefly elaborate the principle behind each of the synthesis techniques and lay down the causes that degrade the naturalness of their speech output.

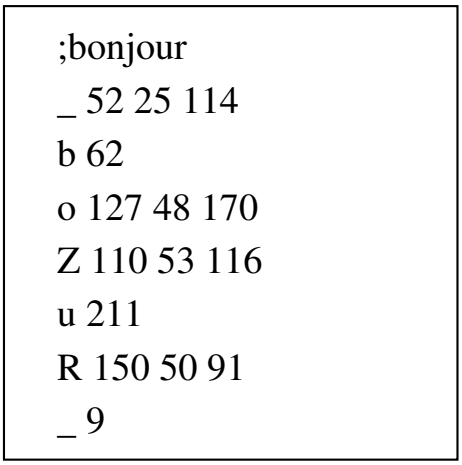


Figure 2.6: The *.pho* file of Mbrola speech synthesis engine (Dutoit et al., 1996).

### 2.3.1 Concatenative Synthesis with Fixed Inventory

The fixed inventory speech synthesis is a speech synthesis that has fixed number of speech units. This size of the fixed inventory is language-dependent; for example, Standard Malay (SM) language has 27 consonant and 6 vowels (Maris, 1980; Teoh, 1994). Therefore, if one want to choose a phone as size of speech unit, the speech database of SM will have 33 number of speech units.

The fixed inventory synthesis model produces a synthetic utterance by modifying the prosodic values of the selected speech units and then concatenates these modified speech units together to form an utterance. The intention to modify the prosodic values of the speech units is to make the synthetic speech sounded natural. For example, lengthening the duration and reducing the pitch values at the end of a synthetic utterance will produce similar speech output to a natural speech at the final phrase of declarative sentence (for SM).

Fig.2.6, taken from Dutoit et al. (1996) shows an example of two prosodic values, pitch and duration, that are used to modify the phonemes of utterance *bonjour*. In the

example, phoneme /R/ is modified with duration of 150 ms and at 50% of the 150 ms duration, pitch is plotted at 91Hz. This data is an example of *.pho* content (phonemes associates with prosodic features), an input file to Mbrola TTS engine (Dutoit et al., 1996).

After the modification of prosodic features takes place, the concatenation process will ensemble the modified speech units. Since there are going to be segmental (amplitude and/or pitch) mismatch at the concatenation points of the ensemble speech units, smoothing algorithm will be applied to smoothen the mismatch (Dutoit, 2008). By using both of the prosodic modification and the application of smoothing algorithm, this synthesis model will be able to produce natural sounding utterances.

Although the concatenative synthesis technique has indeed achieved a significant natural sounding utterance by avoiding to model human speech production system, it is still not up to the naturalness quality demanded by the industrial market. This is largely due to the mismatched concatenation points at joint speech units. Though smoothing algorithms (i.e Linear Prediction (LP), TD-PSOLA, MBROLA), are applied to smoothen the mismatched concatenation points, yet, these algorithms are still unable to make utterance sounded as natural as human speech. Furthermore, some smoothing algorithms degrade the synthesised utterances further by adding unintended "buzziness" noise (Deketelaere et al., 2001). Not only that, the sound produced is still perceived as neutral and robotic.

Taking natural speech in total, by avoiding the prosodic features modification and smoothing process on the concatenated speech units, gives birth to the so-called corpus-based speech synthesis technique.

### 2.3.2 Corpus-based Speech Synthesis

The objective of corpus-based approach in speech synthesis is no longer looking at signal processing as the major role to create natural sounding synthetic speech, but, rather to use it minimally or to avoid it at all. The key idea of the corpus-based approach is to form synthetic utterance by selecting and concatenating natural speech units from a large corpus.

The idea of using natural unit directly as speech units was initiated by the Advanced Telecommunication Research Institute International (ATR) through the system called ATR-vTalk (Black et al., 2007). The approach was later improved by considering both prosodic and phonetic appropriateness in order to choose the most appropriate speech units and implemented on a TTS system called CHATR (Campbell, 1996). Both of these systems were developed by the ATR group. Later, Hunt and Black (1996) improved the CHATR system by selecting speech units in a framework of state transition network and used Viterbi search to choose the most appropriate speech units. The term *unit selection* was used then and until now, it is the standard approach of corpus-based speech synthesis.

Unit selection is based on weighting the sum of two cost: *target cost* and *concatenation cost*. The target cost is the cost calculated between the target unit and the potential candidate units. The candidate with the least cost is more likely to be chosen because of its lower potential to cause a prosodic mismatch. Among the features that are considered for the target cost are the phonetic context, stress and numerical value of pitch and duration (Dutoit, 2008). In Fig.2.7, target cost is represented by the red line.

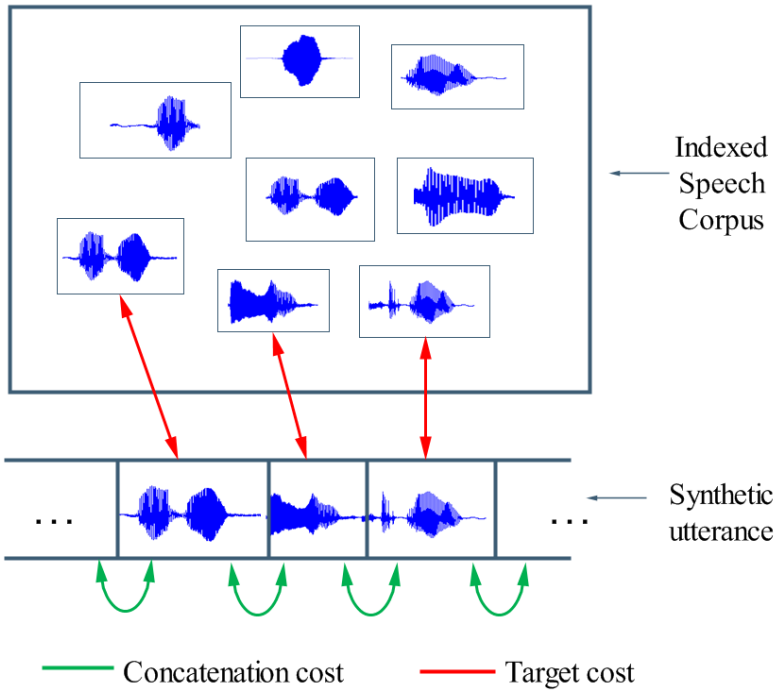


Figure 2.7: Unit selection scheme (adapted from Tokuda et al. (2002)).

The concatenation cost is calculated between two joint candidate units. Speech unit that has the least concatenation cost will be chosen since it is least likely to cause a segmental mismatch. Features that are commonly considered in calculating the concatenation cost are the linear prediction (LP) spectra, line spectrum frequencies (LSFs), Mel-frequency cepstrum coefficients (MFCC) and other frequency-modified spectral representation (Dutoit, 2008). In Fig.2.7, the concatenation is represented by the green line.

Since both of the prosodic and segmental mismatch affect the utterance naturalness, therefore, the most appropriate candidate unit is chosen if the sum of both target cost and joint cost is minimal.

In CHATR system (Campbell, 1996), the synthesis unit is a phone and no

signal modification applied on the concatenation points. New sound is produced by re-sequencing the phones units. Later, the AT&T group developed a TTS system, named the NextGen, by changing CHATRS synthesis units into diphones and concatenating the synthesis units either by just re-sequencing the synthesis unit or applying smoothing technique (HNM and PSOLA) (Beutnagel et al., 1999). The AT&T group later found out that the naturalness in produced utterance was higher when diphones was used as synthesis unit and without applying any signal processing (Beutnagel et al., 1999, 1998).

The unit selection was further improved by changing the fixed unit size into a non-uniform size in Stylianou (2001). The non-uniform size of speech unit managed to reduce the degradation caused by prosodic modification of signal processing technique and the speech signal modeling (Stylianou, 2001).

Despite the successfulness of unit selection to be able to produce highly natural speech output, several problems do come along. First, the naturalness is limited by the size of the speech corpus. The naturalness output of unit selection can be distorted if the coverage of the speech corpus is not enough. Insufficient speech corpus coverage eventually causes least appropriate synthesis unit to be chosen for concatenation. Least appropriate synthesis unit distorts the flow of the speech output because of the potential mismatch either at prosodic features or segmental features, or both. Therefore, to get better selection of speech unit, a larger data is needed. However, the size of speech unit corpus can never be enough, (Black, 2002) since language keeps growing and evolving.

Second, only one type of speech style (or character) can be produced and this is

because synthesis technique uses natural unit. For example, if the speech corpus was recorded with a news reading style, the same style will be carried out if implemented in different domain of applications (e.g. conversation system) and this of course sounded unnatural (Black, 2002). Setting up speech corpora with different styles of recording can solve this problem but the speech variation is yet still limited (Black et al., 2007).

To overcome this limitation in unit selection, the current direction of speech synthesis is now diverted to Hidden Markov Model (HMM)-based speech synthesis or also known as statistical parametric speech synthesis (Black et al., 2007).

### 2.3.3 Corpus-based Speech HMM-based Speech Synthesis

HMM-based speech synthesis can synthesise speech in various voice characteristic easily by changing the HMM parameters. In HMM-based speech synthesis, synthesis process is no longer uses natural speech unit directly, but rather trains the parameters from the natural unit and uses the parameters to synthesise speech.

Before the HMM-based synthesis be able to synthesise input text, the speech corpus will be trained into HMMs. In the training process (see Fig.2.8), acoustic parameters, spectral and excitation, are extracted from the speech database and modelled with the context-dependent HMMs (Tokuda et al., 2002; Black et al., 2007). The context-dependent in each of the HMMs is based on the phonetic, prosodic and linguistics (Black et al., 2007).

For the synthesising process, the input text will be converted into a sequence of context-dependent labels. Using the output of training process, each of the

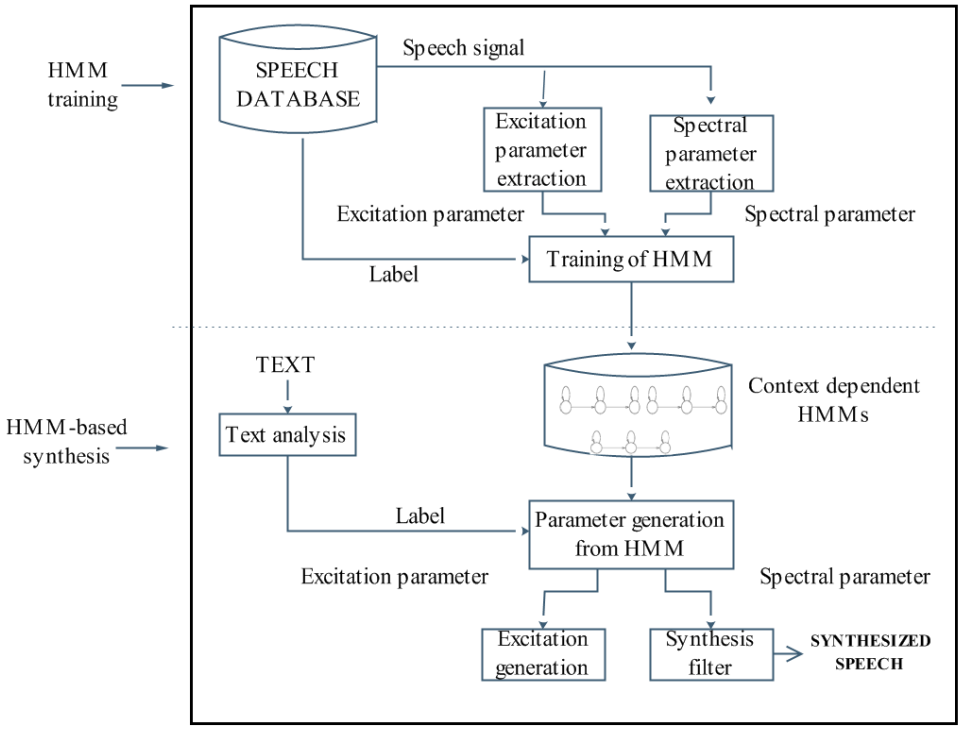


Figure 2.8: HMM-based speech synthesis system (adapted from Black et al. (2007)).

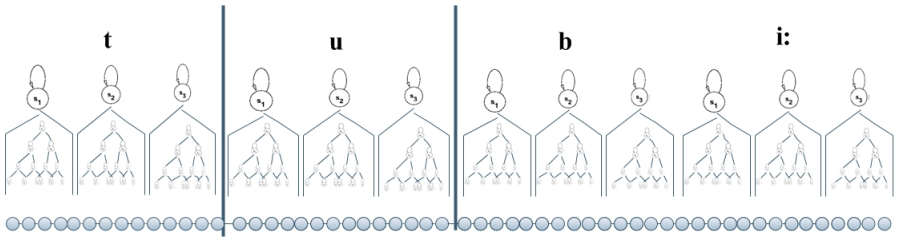


Figure 2.9: The constructed and concatenated HMMs for the words 'to be' (Dutoit, 2008).

sequence labels will have its own context-dependent HMMs and these HMMs will be concatenated. In Fig.2.9, the label phonemes of word 'to be' have their HMMs constructed and concatenated (Dutoit, 2008). Based on the HMMs, an algorithm called speech parameter generalisation will generate the excitation and spectral parameters from the concatenated HMMs and these generated parameters will be used to synthesise a waveform speech for the target word 'to be' (Tokuda et al., 2002; Black et al., 2007).

However, despite the successfulness of the HMM speech synthesis to provide the

control to change the speech characteristics, the naturalness of the speech output is degraded. This is because HMM does not use natural speech unit directly, instead it uses vocoder technique to produce speech. The vocoder technique produces "buzzines" sound while synthesising speech (Tokuda et al., 2002; Black et al., 2007). Besides, the speech quality is also being degraded by the over smoothing algorithm and modelling accuracy (Black et al., 2007).

Looking at the evolution of speech synthesis technique, one can conclude that using natural speech units without applying any signal processing is the technique to produce the most natural sounding of synthetic speech. Thus, we agree with Stöber et al. (1999) that the fast way to satisfy the demand to have a highly natural sounding of speech synthesis is by using a synthesis technique that can avoid or lessen the concatenation points and application of signal processing.

In the next section, we will give an overview of the current Malay TTS state-of-arts.

## 2.4 Overview of SM speech synthesis

Roughly, the research work of SM speech synthesis can be distinguished into three categories based on the speech synthesis techniques they are using:

1. Syllable as synthesis unit.

   Basically, this approach uses syllables as synthesis unit and only one instance existed for each of synthesis unit. During the synthesis process, none or less signal processing is applied between the concatenation points. Since the speech database is a fixed inventory, the same synthesis unit is used in any location

of the word in a sentence and thus, high occurrence happen to both prosodic and segmental mismatch. SM TTS system that was built based on this approach is the Malay Speech Synthesizer (MSS) system developed by Samsudin et al. (2004). Later, Samsudin (2007) proposed to adapt corpus-based approach to improve the MSS to be more natural sounding and re-named the TTS as MSS ver2.

The more sophisticated system that uses the approach of using single instance of syllable as speech unit was proposed by El-Imam and Don (2000).

2. Synthesis with fixed inventory.

   In this approach, diphone is used as synthesis unit in a fixed inventory database and signal processing is used to modify the prosodic values and smooth concatenation points. Fasih$^{tm}$, which was developed by MIMOS, uses this approach. The system uses rule-based approach that predicts the pitch values at phrasal and word levels. The prediction depends on the location of breaks within the target sentence, and then locates the breaks based on Part-of-Speech (POS). The predicted pitch at certain segment of all the target synthesis units will be used as an input to Mbrola engine (Kow, 2005) (see Fig.2.6 as an example of the input file to Mbrola engine).

3. Corpus-based synthesis approach

   Two known research works on SM TTS based on unit selection synthesis technique are the SM TTS system based on Festival framework by Loo et al. (2007) and another one is a TTS system called Malay Text-to-Speech (MTTS) developed by Tan and Shaikh-Salleh (2008) from Universiti Teknologi Malaysia (UTM). Both of these systems are not using prosodic features in selecting speech unit but claimed that the naturalness of speech output generated by unit selection