STUDY ON PHONETIC CONTEXT OF MALAY SYLLABLES TOWARDS THE DEVELOPMENT OF MALAY SPEECH SYNTHESIZER

NUR HANA SAMSUDIN

UNIVERSITI SAINS MALAYSIA

2007

STUDY ON PHONETIC CONTEXT OF MALAY SYLLABLES TOWARDS THE DEVELOPMENT OF MALAY SPEECH SYNTHESIZER

By

NUR HANA SAMSUDIN

Thesis submitted in fulfillment of the requirements for the degree of Master of Science

January 2007

ACKNOWLEDGEMENT

Research will be a very lonesome activity if not for the wonderful researcher of the domain and the environment. Now it has come to the crossing point, I realize I may not arrive to this stage without the experience I gained from people around me.

I would like to thank, my respected supervisor; Dr Tang Enya Kong, for introducing me to the beauty route of research, for the endless motivation and trust, for teaching me to be more patient and wiser. Also to my respected co-supervisor; Dr Chuah Choy Kim, thank you for the support and believe that there is always the best in our self and we can always do better. And also for keep telling me, everything is going to be alright no matter how deadly the road may seem.

It is also an honour to be *acquainted* with very supportive researchers without really actually meeting them in person but they generously providing professional opinion and help online. I must thank *Pascal van Lieshout,* from Oral Dynamic Lab of Graduate Department of Speech Language Pathology, University of Toronto, *Baris Bozkurt* of MBROLA, *Arry Akhmad Arman* of Institut Teknologi Bandung, *Nick Campbell* of ATR, *Naoto Iwahashi* of ATR and *Luri Darmawan* of Komunitas Outsourcing Indonesia. I would also like to express my sincere gratitude for having the privilege of meeting a generous researcher, *Guido Aversano* of ENST, who generously provide information on speech coding and signal modification algorithm.

I am very grateful for the opportunity to do research in Computer Sciences USM which provides a very good research environment. Thank you to the school and Unit Terjemahan Melalui Komputer (UTMK) for providing all equipments that I need. Special thanks to the staff from the school and UTMK research staff: Mr Tan Ewe Hoe, Pn Rohana, Pn Jumiya, Dr Siti Khaotijah, Pn Nour Azimah, Pn Norliza Hani and Pn Noor Azlina who have been very supportive and very helpful throughout my studies here. Also to Fazilah, Fairuz and Fitrah: this thesis will not be completed without your participation! Deepest gratitude to Institute of Graduate Studies and USM for granted me with *Pasca Siswazah Fellowship* for two years.

I want to thank all my dear friends and research colleagues. Especially to Sabrina, Najwa, Lian Tze and Tien Ping for being my shoulder to cry on, for the generous information sharing, constructive arguments and accompany; Adib, for being supportive *big brother* and taking care of our welfares: the postgraduate students. To Sze Ling, Ibraheem, Gan, Sara, Isabelle, Hong Hoe, Hussein, Siew Kin and Pan: with you, life is not just bearable; it is colourful and wonderful as well!

To Fiza, Afifah, Azman, Anuar, Nuha, Amal, Eza, Aisyah, Has, Zurai and Ubay; thank you for being who you are. And thank you for all the support you gave me.

Last, but never the least, to the most important person of my life, my mom and to the members of the family who always giving their unconditional love and support throughout my ups and downs, not only for me as a student but also as their still growing child.

Thank you very much.

Nur Hana Samsudin
2007

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KAJIAN KONTEKS FONETIK SUKU KATA BAHASA MELAYU KE ARAH PEMBANGUNAN PENSINTESIS SEBUTAN BAHASA MELAYU

ABSTRAK

Pensintesis sebutan Bahasa Melayu telah berkembang daripada teknik pensintesis berparameter (pemodelan penyebutan manusia dan pensintesis berdasarkan *formant*) kepada teknik pensintesis tidak berparameter (pensintesis sebutan berdasarkan pencantuman). Kebelakangan ini, teknik pencantuman sebutan makin cenderung menuju ke arah penggunaan korpus atau unit pemilihan pensintesis sebutan. Dalam teknik ini, sebutan yang sudah direkod-awal untuk digunakan dalam pensintesis sebutan, disimpan di dalam korpus sebagaimana ia direkodkan pada asalnya. Maklumat tambahan berkaitan gelombang sebutan juga dimasukkan ke dalam fail bunyi untuk memberikan anotasi yang lengkap pada signal bunyi tersebut.

Walau bagaimanapun, kaedah menganotasi gelombang sebutan kekal sebagai satu kaedah yang tidak standard. Apa yang perlu dianotasi dan bagaimana proses pemilihan unit perlu dilakukan bergantung kepada pembangun pensintesis sebutan itu sendiri dan apakah bahasa yang hendak digunakan serta bidang penggunaan pensintesis sebutan itu sendiri. Ciri-ciri yang digunakan untuk mewakilkan sebutan rekod-awal adalah berbeza dan bergantung kepada bahasa yang hendak dijanakan.

Sehingga tesis ini ditulis, kami masih tidak menemui sebarang kajian yang berkaitan dengan ciri-ciri sebutan yang patut diwujudkan untuk mewakilkan sebutan dalam korpus sebutan bahasa Melayu untuk pensintesis bahasa Melayu.

Oleh itu, tesis ini membincangkan isu bagaimana menghasilkan pensintesis sebutan Bahasa Melayu yang lebih semulajadi. Dalam tesis ini, kami memberikan fokus kepada perwakilan ciri-ciri sebutan peringkat tinggi iaitu konteks fonetik bagi sebutan yang hendak dijanakan. Kami telah melakukan pemerhatian ke atas kesan pemilihan berdasarkan konteks fonetik ini kepada kualiti sebutan sintetik. Hipotesis kami adalah untuk menunjukkan bahawa kita perlu mencari padanan yang paling hampir di antara penyataan sasaran dan penyataan yang direkodkan bagi mendapatkan hasil pencantuman sebutan yang terbaik. Secara hipotesisnya juga, kualiti sebutan sintetik adalah lebih baik jika menggunakan kaedah pemilihan ini berbanding apabila pemilihan dilakukan secara rawak.

Seterusnya, kami juga telah mencadangkan satu templat yang akan membantu sistem pensintesis memilih segmen sebutan yang terbaik untuk menjanakan sebutan sintetik yang lebih baik untuk input teks yang panjang. Ini kerana, kajian kami hanya meliputi aspek fonetik dan kami tidak membincangkan aspek lain yang mungkin mempengaruhi kualiti sebutan sintetik secara terperinci. Templat tersebut akan mencirikan kriteria tambahan yang perlu diambil kira semasa pemilihan unit untuk dicantumkan.

Di akhir penyelidikan ini, kami berjaya memberikan pemeringkatan prestasi dan pemilihan bagi sebutan sintetik yang dijanakan berdasarkan konteks fonetik. Berdasarkan hasil kajian ini, kami mengutarakan teknik lanjutan yang boleh dilakukan untuk meningkatkan kualiti pensintesis sebutan Bahasa Melayu.

STUDY ON PHONETIC CONTEXT OF MALAY SYLLABLES TOWARDS THE DEVELOPMENT OF MALAY SPEECH SYNTHESIZER

ABSTRACT

Speech synthesizer has evolved from parametric speech synthesizer (articulatory and formant synthesizer) to non-parametric synthesizer (concatenative synthesizer). Recently, the concatenative speech synthesizer approach is moving towards corpus-based or unit selection technique. In this approach, the pre-recorded speech segments which are to be used in the synthesizer are stored exactly as how it is recorded. Additional information of the speech waveform is attached to the sound to provide proper annotation of the speech waveform.

However, annotations of the speech waveform remain as a loose standard. What should be annotated and how a unit selection process is carried out rely heavily on the developer, and target language, as well as the target domain of the synthesizer usage. Features used to represent pre-recorded speech are varied and treated as language dependent.

Until this thesis is written, we are still unaware of any study related to what speech features should be made available in a Malay speech corpus for a Malay speech synthesizer.

This thesis addresses the issues of producing a more natural sounding speech synthesizer for Malay. We focus on high level representation of speech features which is the phonetic context of the speech to utter. We conducted an observation on the effect

of phonetic context to the quality of concatenated speech. Our hypothesis is to show that, to get the best concatenative speech result, we have to find similar or closest match of phonetic context, between the recorded utterance and target utterance. Hypothetically also, the output quality of this selective method will be better than when we select a segment in random.

We also proposed a template which will guide the system to select the best candidate of the speech segment to produce a better synthesized speech for longer utterance. This is because, our study covers only the phonetic aspect of the speech and we did not discussed on other aspects of speech in detail. The template will detail out additional criterion which need to be followed during selection of unit to be concatenated.

At the end of the research, we are able to give out the performance and preference rating of the concatenated speech which is based on phonetic context. Finally, we presented the future work to further improve Malay speech synthesizer.

CHAPTER 1
INTRODUCTION

Speech is one of the many ways by which human communicate. Speaking is a process which we carry out without full realization. However, it consists of a very complex process, starting from the inhalation of air which moves from the lungs to the vocal cords and vocal apparatus to create the linguistic acts in the form of language that communicate information from an initiator to recipient (Wikipedia, 2002a).

To synthesize speech using computer is not the same as human to speak. Different approaches may be used. For instance, there is a technique of producing speech using a physical model of speech production that includes human articulators. It is also known as articulatory synthesis. There is also a method which model human vocal tract and folds using electrical devices and manipulate them based on varying the formant frequency (or pitch) *viz* formant synthesis. A more recent approach uses pre-recorded speech to produce novel utterance. This approach is named as concatenative synthesis.

1.1 General Speech Synthesis Architecture

Speech synthesis is the process involve to produce speech by machine (computer in our domain) using a few speech chunks. In this thesis, we will discuss the technique of concatenating speech segment to produce totally novel utterance.

The general architecture of a speech synthesis is shown as Figure 1.1. The input to the speech synthesis system could be in the form of raw text or any form of desired tagged text. Text analysis module will fully transform the input text into readable form for further processing. In this way, the module will change any symbol or number and also abbreviation into fully written alphabet characters. Phonetic analysis will take care of how each word should be pronounced by representing the input text into phonetic form. It will also handle homograph disambiguation issue. Prosodic analysis is the module where the system will analyze and design the intonation of the synthetic speech to make it sounds more natural, ie as produced by a human. And finally, speech synthesizer module will gather all information gained from previous modules to render the speech.

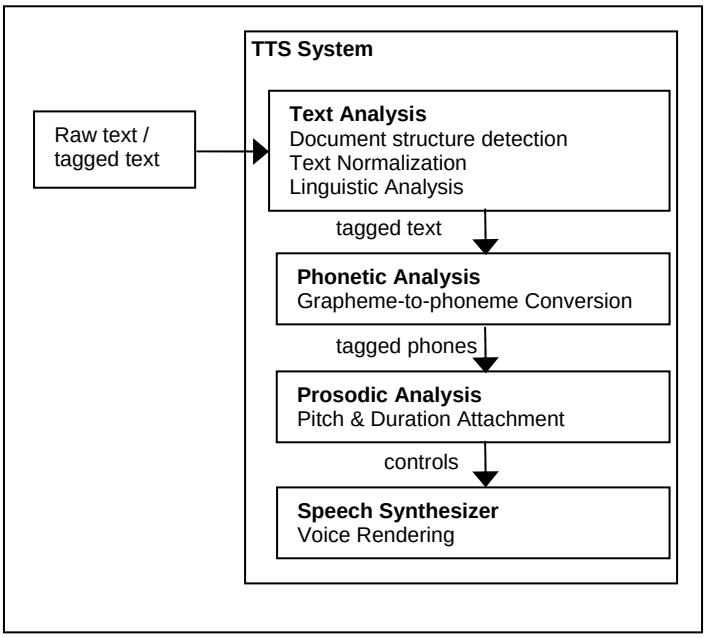

Figure 1.1: General Text-to-Speech Architecture (modified from Huang et al., 2001)

Our scope of research will give special attention to phonetic analysis and speech synthesizer module. We will explain deeper on the scope of our research in section 1.4.

1.2 Speech Synthesis Usage

Speech synthesis technology not only serves to help blind people and others with disability. It could also make everyday activities simpler and could also make the process

of learning more interesting. For example, speech synthesis is used in the software for the blind (*screen reader*) which the software will read out aloud based on the movement of user's mouse. Stephen Hawkings's (a great physics scientist) illness also leads to the need of speech synthesis[1] system. Speech synthesis allows one to collect contents of a reading material while doing other tasks. It could also be used in *edutainment* like adding speech features in courseware. Other various usages of speech like telecommunications services purposes, talking toys and books and part of man-machine communications.

1.3 Current Issues in Speech Synthesis System

For each module in speech synthesis architecture, a lot of deficiencies have been identified. It may be categorized as language dependent and slightly language independent. For certain modules, issues and solution to overcome the problems are language dependent. For example: issues related with text analysis and phonetic analysis modules are:

- text analysis
    - text normalization

      Handle the conversion process from non-orthographic text into common orthographic transcription
    - linguistic analysis
- phonetic analysis
    - homograph disambiguation

      Disambiguate words with different senses to determine proper phonetic representation (how it should be pronounced)

[1] Refer to: http://www.hawking.org.uk/disable/dindex.html

- morphological analysis

  Analyze the morpheme component in word to attain the correct phonetic pronunciation

- letter-to-sound conversion

  Includes general letter-to-sound rules to produce accurate pronunciation

Some process is language independent and some are adaptable from one language to another like prosody analysis module which determines the intonation contour of the synthesized speech. The module could be adaptable for similar type of language, particularly for unstressed and non-tonal language. Same goes for how we could concatenate a few speech chunks to produce a new utterance.

However, there are more deficiencies regarding the quality and also the size of the speech chunk required to have a complete synthesizer. Among them are:

- Prosodic analysis module
  - Stereotype of synthetic speech

    By pre-determining how the intonation contour should be, a stereotype speech synthesis will be produced
- Speech Synthesizer
  - Distorted speech

    Speech produced by synthesizer in general is noticeably distorted.

    Scientists in speech processing domain are still looking into how speech synthesis quality could be further improve so that it sounds more human

- Expressive speech

  An extensive study needs to be conducted to produced a synthesizer which able to include the emotion of the speech.

- Very big speech corpus size

  Speech synthesis system may require a big speech corpus to enable the system to produce any desired utterance. However, the issue of space required to store all pre-recorded speech is always a major disadvantage to create a complete speech corpus.

There could be other issues in the current speech synthesis evolvement. In this thesis, we will focus on handling distortion in speech synthesis.

1.4 Research Contribution

The study of distortion in synthesized speech is still an ongoing research interest. One of the methods is by using real time speech segment selection. However, we first need a complete speech corpus. To develop a complete corpus for Malay speech synthesis, we need to identify the information required to put into the speech corpus. This thesis will discuss information which possibly could be added as one of the corpus information. We will study in detail on the affect of phonetic context in Malay, specifically on the adjacency and the position of the syllable. We select the syllable as our basic unit because syllable is able to retain the prosodic aspect of the speech better than smaller unit. This research mainly study Malay pronunciation affect on the phonetic context at the word level for the construction of Malay speech corpus for unit selection speech synthesis.

At the end of our research we will be able to determine whether phonetic context may influence degradation or improvement of synthesized speech quality. It is important as one of the key to construct the design of Malay speech model.

Also, we will propose a layout to synthesize a complete sentence. Since the focus of the study is to improve naturalness at the word level, the quality will not be very satisfying if we solely used the proposed approach to produce a synthesized sentence. By providing the layout in form of template, we believe the quality of synthesize speech will improve significantly.

At the time of writing, we are not aware of any study on the construction of unit selection speech corpus for Malay text-to-speech. We would like to consider this study as a novel contribution in Malay TTS study.

1.5 Some Terms and Concepts

1.5.1 Type of Unit

A pre-requisite of speech synthesizer is a corpus. A corpus contains a collection of pre-recorded speech and properly annotated to clearly represent each unit. A few examples are shown below:

Table 1.1: Type of unit and corresponding segment of each unit for the word "terjemahan" (translation)

Type of unit	Label of corresponding waveform
word	tərʤəmahan
demi-syllable	_tər + tərʤə + ʤəma + mahan + han_
syllable	tər + ʤə + ma + han
diphone	_t + tə + ər + rʤ + ʤə + əm + ma + ah + ha + an + n_
phone	t + ə + r + ʤ + ə + m + a + h + a + n
half-phone	$t_L + t_R + ə_L + ə_R + r_L + r_R + ʤ_L + ʤ_R + ə_L + ə_R + m_L + m_R + a_L + a_R + h_L + h_R + a_L + a_R + n_L + n_R$

* superscript L and R represent left and right half-phone

1.5.2 Prosodic Features

There are a lot of definitions on prosody. We select Dutoit's (1997) definition which summarizes most prosody definitions.

> "The term prosody refers to certain properties of the speech signal such as audible changes in pitch, loudness and syllable length. For some authors the set of prosodic features also include (other) aspects related to speech timing such as rhythm and speech rate."
>
> (*ibid*: 129)

Some even call intonation as a synonym to prosody (*ibid*). However, prosody is measurable while intonation is subjective to human perceptibility. Thus, prosodic features will directly refer to three values in speech: pitch, intensity and duration.

1.5.2.1 Pitch

Pitch representing the frequency of sounds. Frequency of a sound refers to the number of complete cycles in a second. However, this term is used to describe simple periodic waveforms. For a complex periodic waveform like speech, fundamental frequency or F0 is the correct name to give (Huckvale, 2000).

1.5.2.2 Intensity

Intensity can be defined as the average amount of energy passing through a unit area per unit of time in a specified direction (OMP, 2003). According to Gotfrit et al. (1995), acoustic envelope is a characteristic variation in the sounds overall amplitude from the moment a sound begins until it ceases. There are relations among intensity, energy and amplitude which are corresponding to the acoustic representation of loudness (Dutoit, 1997: 130). However for this research scope, we focus only on the overall intensity contour. For brief description on intensity contour, see Huber & Runstein (1992).

1.5.2.3 Duration

Duration refers to the length of pronouncing a sound, whether a syllable, a word or even a phoneme. For tonal language, different duration segment may signify different meaning.

1.5.3 Phonetic Context and Transcription

Different sequences of phonemes within a word or between two consecutive words create difference in pronunciation. This could be influenced by phonetic sequence of the text or we call it as phonetic context. Phonetic context also covers issue on position of segments, manner of production and a few others. In short, phonetic study could cover anything related to the construction of the word or sentence, depending on the scope of study.

There is another similar but different term: phonemic. There is a fine line differentiating phonemic and phonetic transcription.

Table 1.2: Difference between orthographic, phonemic and phonetic transcription

Orthographic Transcription	Phonemic Transcription	Phonetic Transcription
<gelak> (laugh)	/gəlak /	[gəlaʔ]
<buah> (fruit)	/buah/	[buwah]
<siang> (day)	/siaŋ/	[sijaŋ]
<taat> (obey)	/taat/	[taʔat]

We can see that phonemic transcription mapped each grapheme to corresponding phoneme. Phonetic transcription on the other hand is an advanced form of phoneme sequence that reflects how a word should really pronounce despite how it is spelled. It is influenced by the original sequence of phoneme, the language and also the place and manner of articulation of sequence phoneme.

1.5.4 Instances in Unit Selection

We will frequently use *instance* to represent a unit in speech database. In unit selection approach, there is a lot of similar unit name and label but with different features value where one is proper to be selected than the other depending on the situation and adjacent unit. So instance will be use to differentiate between the same label but different features value.

1.5.5 Corpus and Database

In this thesis, we will make a slight distinction between these two terms although both referring to the speech collection used in speech synthesis. We will use the term database when the storage of speech we referring to are a type which the sound has been properly extracted from its original recording. We use the term corpus when we referring to original speech storage without any modification but has been aligned to its corresponding phonetic representation. However both will have their corresponding annotation of features.

1.5.6 Syllables

Syllable structure can be view as in Huang et al. (2001: pp 52). We find that the definition of syllable below is very precise to describe our usage of the term:

> Syllable is a phonological structure composed of speech sounds. Words are made up of syllables. The syllable is the domain of association for such phenomena as accent, stress and lexical tone. Syllables are generally considered to be composed at a number of constituents: onset, rhyme, nucleus and coda.
>
> (Maidment et al., 2006)

1.6 Thesis Outline

This thesis is organized into 7 chapters. This chapter gives a brief introduction on speech synthesis such as the capability of a speech synthesis, the architecture of the whole framework and also issues related to deficiency of speech synthesis that scientists face.

In Chapter 2, we will give a background survey on concatenative speech synthesis. Readers may skip this chapter if readers are very familiar with speech synthesis domain. We will discuss in detail the approaches in concatenative speech synthesizer.

In Chapter 3, we will have a literature study on the existing speech synthesizer in proposed approach. We will highlight current issues in speech synthesis and the difference contribution of the evolving work in this chapter. We will also present the unit selection speech synthesis framework and how our study can fit into it.

Chapter 4 reports our study of the affect of phonetic context on pronunciation in Malay at the word level. This study is important to further carried out on the construction of Malay speech corpus for unit selection speech synthesis.

In Chapter 5, we will describe how we implemented the proposed idea inside our prototype to deliver a synthesized speech. We also highlighted the architectural differences between two synthesizer approaches. The output of both approach are also attached in the disc.

In Chapter 6, we evaluated the quality of synthesized speech between the proposed method and another synthesizer which is used as a comparison. We will discuss the results of the evaluation and the rational behind the respondents rating.

Finally, Chapter 7 will conclude our work and we will also talk about how this study can be further improved in the future.

We also added some terms and terminology which may use through out this thesis regularly but treated as a common terms and terminology. They are attached in Appendix L.

CHAPTER 2
BACKGROUND STUDY

In general, there are three types of synthesizer: articulatory synthesis, formant synthesis (source-filter) and concatenative synthesis.

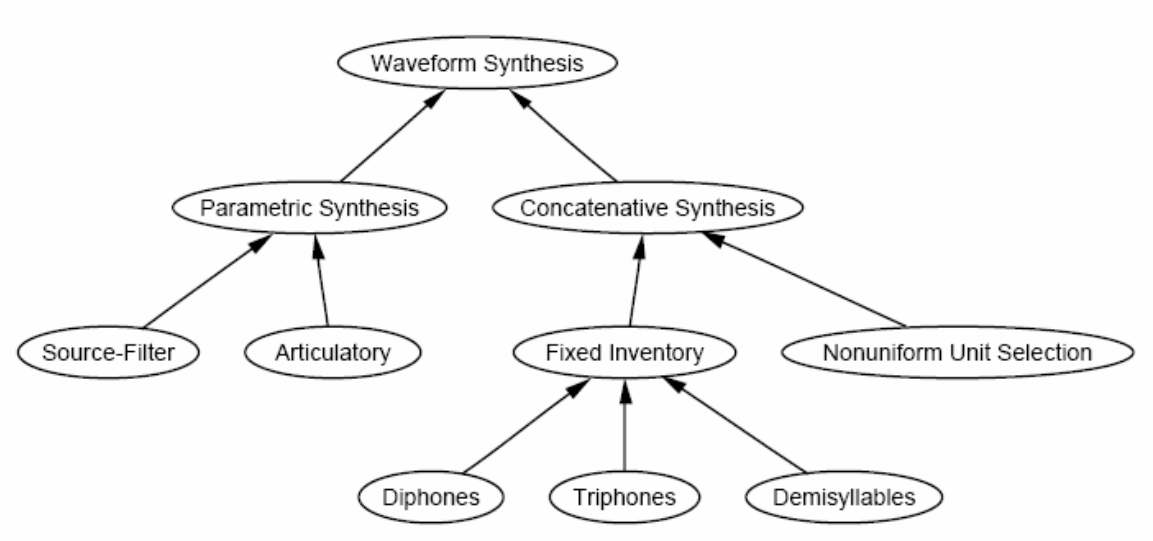

Figure 2.1: Classes of Waveform Synthesis Methods (Schwarz, 2004)

Schwarz (2004) classified synthesizer starting from waveform synthesis. Waveform synthesis can be divided into two: parametric synthesis and concatenative synthesis. Articulatory and formant synthesis are under parametric synthesis category. Concatenative synthesis may be formed by fixed unit representation or a non-uniform unit representation.

In a fixed unit representation, usually the inventory[2] of speech is made up by one specific type of segment. And only one speech segment is available for each phonetic correspondence or what we call as instance.

In non-uniform unit selection approach, more than one candidate is available before selection of the most optimal candidate is made. These multiple instances will have a set of information (also known as features) which differentiates between instances. Unit selection which allows multiple labelling make it possible for a speech corpus to have variety type of unit; e.g.: a speech corpus which has diphone and triphone labelling. Contrary to fixed inventory approach, which only has one type of a pre-recorded segment is stored inside the corpus, the unit present in speech corpus of non-uniform unit selection inventory may be as varied as the developer wish (depends on labelling[3]).

A few examples on two classes of concatenative speech synthesis are shown below:

Table 2.1: Summary of Approach Used in Available Product/Research Centre

Speech Class	Product/Research Centre
Fixed Inventory	MBROLA (diphone), Festival, Microsoft SAPI (until ver. 5.0)
Non-Uniform Unit Selection	CHATR, AT&T Next-Gen TTS System, Festival2, Nu-MBROLA

We refer to the fix inventory approach as conventional speech synthesis since it emerges before non-uniform unit selection approach.

2.1. Conventional Speech Synthesis

Conventional speech synthesis required all existing segments (for the particular language) to be available inside their pre-recorded speech segments. As stated at the beginning of this chapter,

[2] which consist of pre-recorded speech chunk (a.k.a. speech chunk)
[3] Labelling and aligning of speech to the corresponding phonetic and acoustic features are also known as annotation

only one instance is available for each speech segment the targeted language. The recording will follow certain preset values (the pitch, intensity and duration range for example) and these values will be documented as a record for further reference. During concatenation, each corresponding segment will be selected and join together consecutively according to the input sequence. To minimize perceptual distortion, a few algorithms to smooth the point of concatenation and to increase the naturalness of the speech have been introduced. This process may be shown as Figure 2.2 below:

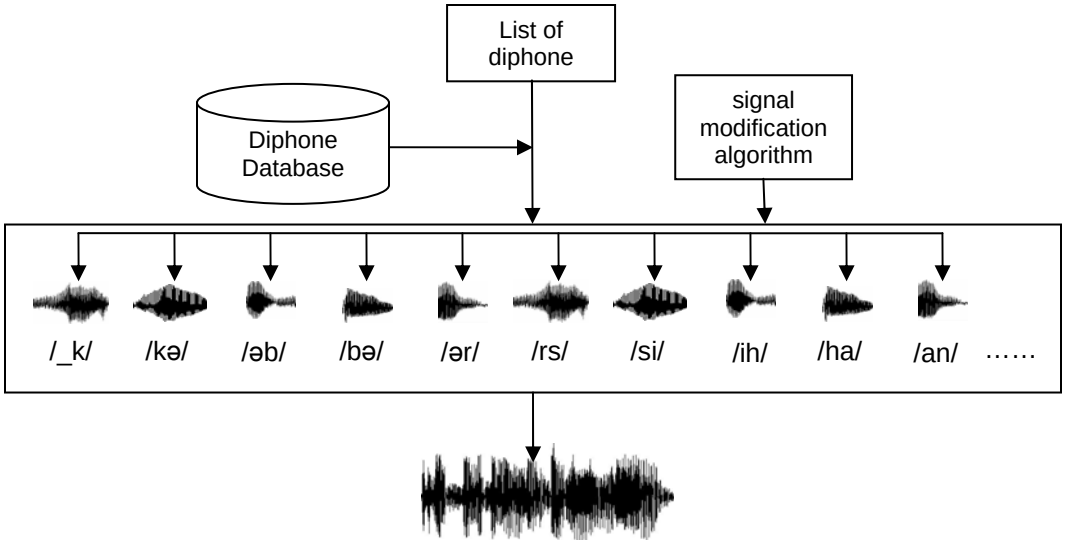

Figure 2.2: Process in conventional speech synthesis system for diphone approach

From this figure, diphone database will provide the segments (in the diagram the segment is diphone) requested based on list of diphone which is generated based on the input text. Assuming the input is a phrase: “Kebersihan tanggungjawab bersama”. Thus the database of diphone will provide all diphone segments to produce the desired utterance.

2.2. Unit Selection Speech Synthesis

In general, unit selection speech synthesis uses a very large recorded speech corpus (more than one hour recording) with corresponding annotation. However, instead of having the speech segment extracted and stored in isolation, unit selection enables the engine to store the pre-

recorded speech exactly as how it is recorded and extracted them during the runtime (this method is known as online synthesizer). To enable further manipulation, annotations of waveform segment need to be performed.

In unit selection speech synthesis, annotation for the pre-recorded speech is very crucial since it represents speech information for a segment to be selected for concatenation. For each instance, it will be aligned with their corresponding prosodic value or/and phonetic context of the speech segment. Beside that, other linguistic or acoustic information might also be attached. We will discuss unit selection system architecture in the section after next.

2.3. Comparison between Conventional and Unit Selection Synthesizer Approach

The distinctive differences between unit selection speech synthesis and conventional synthesis is how each approach represents their pre-recorded speech and the level of involvement of signal modification to minimized perceptual distortion of the concatenated speech.

To handle distortion at the concatenation points (for conventional approach), a lot of algorithms has been introduced. Among them are: PSOLA, MBROLA and LPC (Dutoit, 1997; Conkie, 1999; Huang et al., 2001). That is what has been implemented in conventional speech synthesis.

Unit selection concept on the other hand, tries to minimize; if not able to avoid; wave modification by providing more instances to choose to be concatenated. It is thus possible to select a set of candidates with almost perfect combination which requires a small signal modification to smoothen the speech.

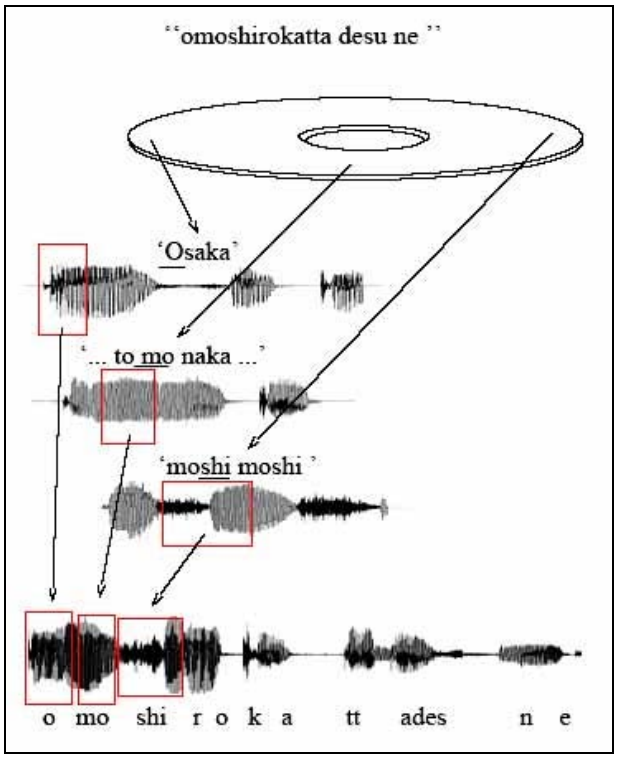

Figure 2.3: CHATR's Unit Selection Approach (Campbell, 1997b)

For example, Figure 2.3 above shows how CHATR's make the selection. During selection, only the closest value in index of pre-recorded speech (the annotation scheme) to the target unit segment will be extracted and concatenated to produce novel utterance.

Annotation of the corpus is relying on what are the features of the speech the developer want to store and use to produce the desired speech. The design of the features to be presented is also depends on what is the target language one wants to generate at the end. In conventional approach however, the information of pre-recorded segment is prepared and stored to enable the system to manipulate the signal to produce the desired speech.

2.4. General Architecture of Unit Selection Speech Synthesis

Although unit selection synthesis architecture may not have a huge variance with conventional speech synthesis, we would like to highlight on some differences in their architecture. This

architecture present roughly on unit selection approach. We will also detail out each component to present a clearer picture on how the components work.

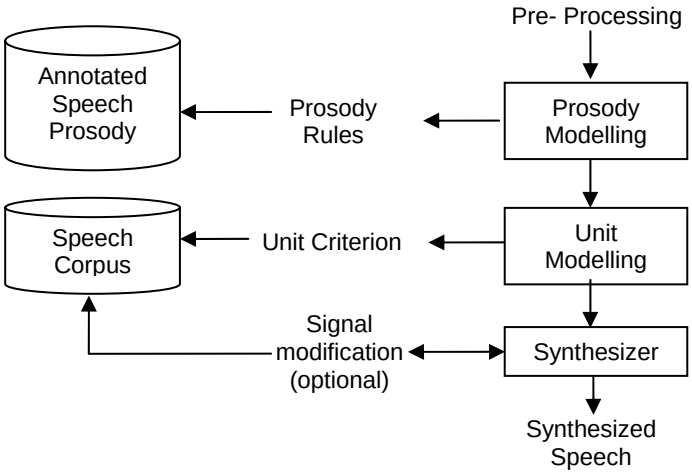

Figure 2.4: General Component in Unit Selection Synthesizer

Figure 2.4 above shows roughly what is inside unit selection synthesizer architecture. This architecture is not absolute. The input of the component shown above is triggered by the output from previous modules. Prosody modelling here refers to the process of the assigning prosody value based on the prosody structure in the pre-recorded speech. It could also be predicted from either rule generated or corpus-learnt. It is not, however, assigned value as what prosodic analysis does in conventional speech synthesizer.

Unit modelling here refers to the searching and selection of candidate to be concatenated. There are varieties of approaches. It can be categorized into *single level* selection or *multi-level* selection. In single level selection, the speech corpus is segmented into a uniform representation which applies to the whole speech corpus; ie: if diphone is a unit representation, the diphone available in the speech corpus is annotated. The multi-level selection on the other hand has more than a type of unit to represent a wave file. This will provide flexibility of unit to be concatenated as compared to having phrase or word segmentation only. The method are usually designed in such a way the selection process will automatically select the longest

sequence of unit which match with the target utterance from the speech corpus. For example: a wave file in a corpus might be labelled correspond to the phrases available in the wave file. For each phrase, there also consist of word, syllable, diphone and phone segmentation. Both example of existing system will be presented in Section 2.5 after this. The main purpose for multi-level selection is to pick the longest available recorded utterance from the corpus to be concatenated to retain the naturalness of speech in the recording.

Unit criterion in the figure also highlighting on the possible parameter or features for selection. This may be set based on the speech model of the target language as what we going to discuss in Chapter 3. It is also important to highlight the method for selection of unit mostly will be based on calculation as shown in Figure 2.5 below:

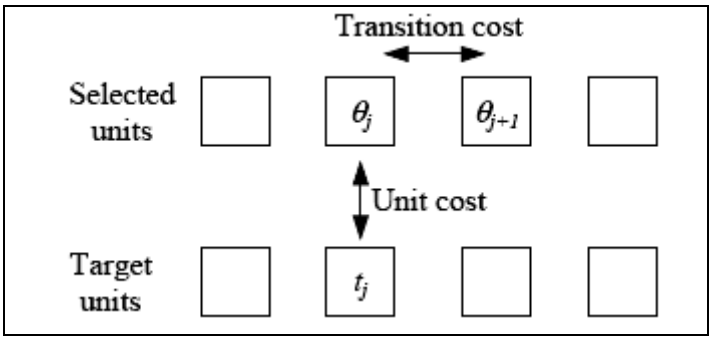

Figure 2.5: Target Cost and Concatenation Cost is the two costs which need to be determined to get the most optimum unit (figure from Huang et al., 2000)

Figure 2.5 highlighting two type of cost which is used to calculate the best available segments that have to be selected: transition and unit cost. The transition cost is to measure the distance of spectral differences between two sequence of units to ensure there are well joined, while unit cost referring to measuring the distance between selected segments with the target or pre-determine segment to be concatenated. Both of this cost is called as cost function. It will measure distortion which involving both cost.

Assuming Θ is the transition cost and T is the unit cost, we will obtain the distortion occurrences for the target unit by summation of both costs:

$$d(\Theta,T)=\sum_{j=1}^{N} d_u(\theta_j,T)+\sum_{j=1}^{N-1} d_t(\theta_j,\theta_{j+1})$$

Equation 2-1: Cost Calculation

Where d, refers to distortion and θ, refers to a speech segment.

Thus, $d_u(\theta_j,T)$ is unit cost of using speech segment θ_j within target T.

And $d_t(\theta_j, \theta_{j+1})$ is the transition cost of concatenating speech segment θ_j and θ_{j+1}.

The smallest cost value means the best sequence of segment which is the best to select from.

Speech corpus may be referring to the same corpus as annotated speech prosody (if we want). The final module is the synthesizer where the selected segments will be put into sequence and then concatenated to form a new utterance. Signal modification may still be used. Among the most popular approach are the Synchronous Overlap and Add (SOLA), Pitch-Synchronous Overlap and Add (PSOLA) and Harmonic plus Noise Model (HNM).

We will see in the next section the detail architecture of unit selection speech synthesizer based on their individual approach.

2.5. Overview on Unit Selection Speech Synthesizer

We presented overview on conventional synthesizer in Appendix A. For unit selection synthesizer, we present it here. We selected 4 existing unit selection speech synthesis systems architecture.

2.5.1. CHATR

The CHATR speech synthesis system was developed by Department 2 (Prosody Interpretation and Speech Synthesis) of the ATR Interpreting Telecommunications Research Laboratories,

Japan. CHATR synthesizer uses unit selection approach which is based on a technique called Re-Sequencing System of Unit Selection (Campbell, 1992a; 1996). Concatenation is implemented by using a re-sequencing of carefully selected phone-sized segment from a pre-recorded speech corpus (Campbell, 1996). The difference of their idea from conventional synthesizer (and a few other unit selection approaches) is that there is no need for signal processing to smooth the concatenation points, beside the unique corpus design. CHATR system relies on the external source of its speech corpus (see Figure 2.6) and reproduces novel utterances using carefully selected segment of the recorded speech of this external source.

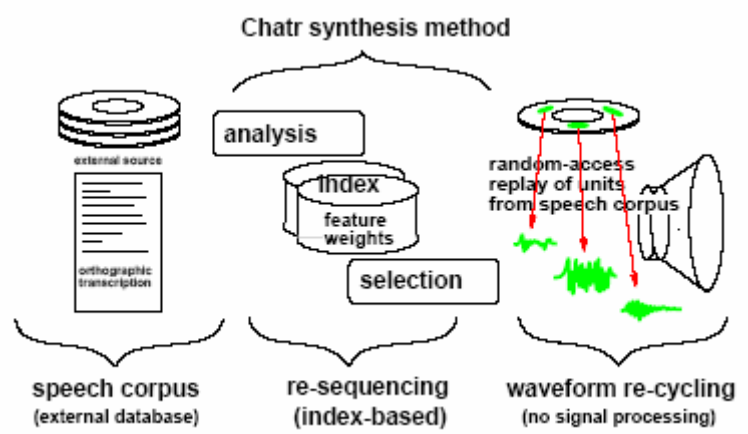

Figure 2.6: CHATR's speech data outside the synthesizer (Campbell, 1997b)

There are a few additional processes in CHATR's re-sequencing synthesizer. One, it needs an index of phones prepared together with the prosodic characteristics for each utterance of the speech corpus. The re-sequencing approach is functioned to determine an optimal sequence of unit to be replayed from original speech to give the best estimation to the desired utterance from the segments available in a given speech corpus. Thus CHATR will analyze the corpus to allow the engine to make prediction on the aspect of prosody in parallel. And the third thing needs to be prepared is the selection mechanism. This is because, there is possibly more than one instance can be use to form an utterance, thus the system must be able to select the best candidate available. Unit to be selected may be based on the annotation value of the speech corpus or certain prediction value which both are pre-determined rules.

In CHATR, the synthesis method is a language and speaker independent (CHATR, 1997). Meaning, CHATR can depend on their speech corpus with the corpus information to produce desired utterance. And CHATR also isolate each speaker's speech from one and another. This is important due to the fact that the online unit selection will retain the prosodic criteria of the original speech inside the novel utterance. Hence, it is very important to provide sufficient data (speech information) so that speech in the corpus is able to produce the input into speech as human as possible (CHATR, 1997).

CHATR maintain the pre-processing of the input text: text analysis and phonetic analysis (refer Figure 1.1). CHATR implementing re-sequencing of unit based on indexing corpus. This is what has been pointed out as significant difference by Campbell (1996; 1997a; 1997b).

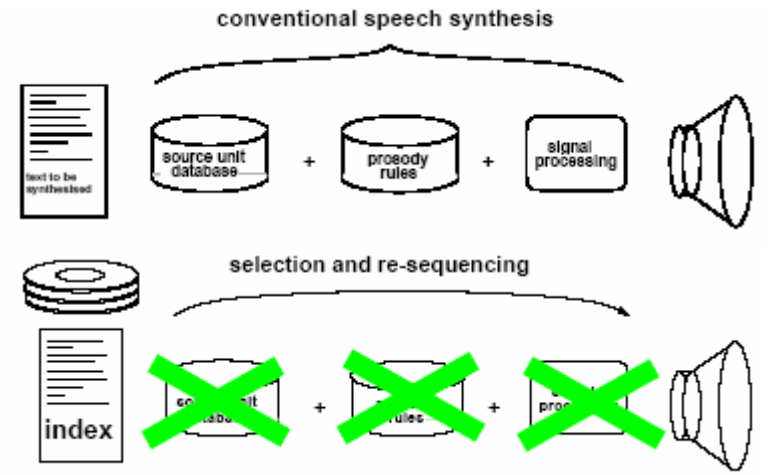

Figure 2.7: Comparison between CHATR's re-sequencing technique and conventional synthesizer (Campbell, 1997b)

Based on Figure 2.7 we can see that CHATR do not have the 3 components: unit database (pre-recorded speech segments), prosody rules and signal processing (waveform modification). CHATR substitute them with their indexes speech corpus and applied an algorithm to select the

best unit to pick for concatenation and simply concatenate the sound without wave modification (see Figure 2.8).

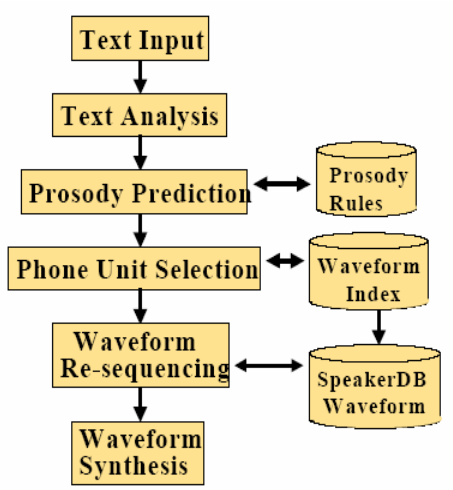

Figure 2.8: Speech Synthesis Method of CHATR (Iida et al., 2000)

As the result, we can see the architecture of modules in CHATR is like the above figure. Text analysis module is still needed before further processing. CHATR will model the intonation based on prepared prosody rules. It is suffice to state that contrary to conventional prosodic rules which emphasize on manipulation of prosody values, the prosody rules here referring to the context of prosody modelling which is extracted (or learned) out from the pre-recorded speech corpus. These intonations rules are templates retrieved form their pre-recorded speech corpus (CHATR, 1997) and preset model. Phone unit selection module will select the closest candidate to the target calculation. Finally waveform re-sequencing will concatenate the entire selected unit to form comprehensible synthesized speech.

2.5.2. AT&T

AT&T Next-Generation TTS System is developed by AT&T Lab Research. The AT&T Next-Generation TTS system is a hybrid of the previous Flextalk system by AT&T, the CHATR system by ATR and Festival system from University of Edinburgh (Beutnagel et al., 1999a; Beutnagel et al., 1999b; Conkie et al., 2000).

Flextalk provide modules for text analysis and phonetic/prosodic specification (Beutnagel et al., 1999a; Beutnagel et al., 1999b; Conkie et al., 2000). Unit selection is based on CHATR's implementation with extensive modification. Contrary to CHATR, AT&T's synthesizer allows signal modification. Typically AT&T uses Harmonic plus Noise Model (HNM). But it also allows some flexibility in the system so that the system is also able to choose another prosody modification algorithm like PSOLA or no modification at all (Beutnagel et al., 1999b; Conkie et al., 2000).

AT&T Next-Generation TTS System's architecture can be picture as figure below:

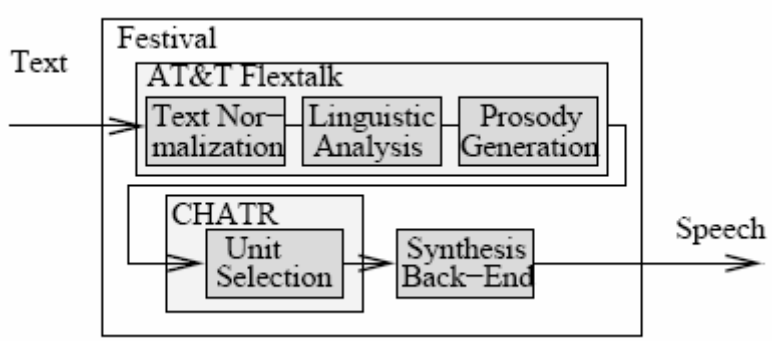

Figure 2.9: AT&T Next-Gen TTS system architecture (Beutnagel et al., 1999b)

One significant difference between the AT&T synthesizer and CHATR is the basic unit use to annotate/segmenting the speech corpus. CHATR uses phone as basic unit. AT&T on the other hand, uses half-phone. This is because they found that it is possible to produce natural sounding synthesized speech by using phone; however the quality is often inconsistent (Conkie, 1999). Thus, AT&T Next-Generation TTS System develop a unit selection and synthesis algorithm that allows finer control than CHATR system by applying selective prosody modification and implementing finer control over unit that is chose for synthesizer (Conkie, 1999).

2.5.3. Multisyn (Festival 2)

In previous section, we did mention Festival by CSTR as one of the popular framework and modules used by researchers to produce speech synthesis system. Originally, Festival uses diphone concatenation (Clark et al., 2004; Black & Taylor, 1997). Later they focussing on limited domain speech synthesis and currently, they are changing their direction to general purpose unit selection engine which they named: Multisyn. It is built in Festival framework and using Festival provided tools.

The Multisyn algorithm works by predicting target utterance structure (Hofer, 2004) from input text. It will be followed by pre-selection and concatenation of the best candidate sequence found. The process of synthesizing in Multisyn can be summarized like Figure 2.10 below:

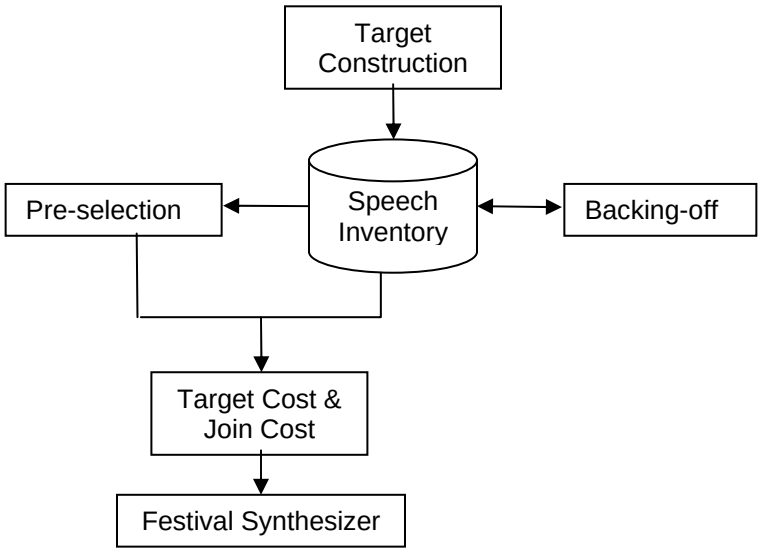

Figure 2.10: Illustrated based on Clark et al. (2004) and Hofer (2004)

Festival 2 also uses diphone as smallest unit representation. Among the rational are phone are extremely difficult to join (Clark et al., 2004). Although diphone may cause difficulty to ensure a full coverage of segments and half-phone seems like able to handle the problems of joins boundaries; half-phone process will be twice longer than diphone. Thus, to make a rapid system

development, they select diphone as unit representation. They also not considering bigger unit than diphone for a time being.

Multisyn's module of unit selection algorithm (Figure 2.10) can be described as below (Clark et al., 2004; Hofer, 2004):

- target construction
    - target utterance structure is predicted from text
    - only phrasing and pronunciation are predicted. Prosodic aspect is omitted
    - a sequence of phones with an appropriate linguistic structure is produced

- pre-selection
    - target phone sequence is converted into target unit sequence (diphone)
    - a list of candidates for each of the units is constructed from the database
    - the list contains all diphones of the target type in the database.

- backing-off
    - this module will be call when the inventory does not contain a specific diphone listed from pre-selection module (missing diphone)
    - diphone will be substitute with other closely related diphone based on predefine rules. Eg: close vowel will be substitute with mid vowel
    - Since it is difficult to obtain a suitable substitute phone sequence, worst case scenario allow any substitution although it means there will be a mismatch in phone sequence either with the diphone preceding or following the substitutional diphone