PHONE-BASED SPEECH SYNTHESIS USING NEURAL NETWORK WITH ARTICULATORY CONTROL

By

LO WAI KIT

盧偉傑

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF PHILOSOPHY

DIVISION OF ELECTRONIC ENGINEERING

THE CHINESE UNIVERSITY OF HONG KONG

19TH AUGUST, 1996

Acknowledgement

The author would like to take this opportunity to express his greatest gratitude to the Croucher Foundation for her generous offer of a studentship over the past two years. Without her financial support, completion of this work will be impossible.

In addition, I would also like to thank Prof. P. C. Ching for his supervision and discussion that leads to the success of this work. I would also thank Dr. T. Lee for the discussion and comment, Mr. Francis Chik for helping me cutting and labeling the speech data for inter-syllable pause properties, my peers in the Signal Processing Laboratory of CUHK, Mr. S. K. Ip, Mr. W. K. Lai, Miss X. X. Niu, Mr. William Pang, Dr. H. C. So, Miss L. Tang, Dr. S. Q. Wu and Mr. Z. L Yu for their valuable advice and assistance. Thanks are also due to Mr. M. H. Ko, Mr. W. F. Liong and Mr. K. O. Luk for their technical assistance and help offered.

Last but not least, I am also indebted to my parents, my whole family for their encouragement and support which is essential for the completion of this work.

Abstract

Speech synthesis has been an active research topic for a long time. It has been applied to many practical areas. It is employed for both enhancing existing tools (such as machine interfaces) and providing speech aids where its application is a must. However, as application requirement is increasing, there is an ever growing demand of more advanced synthesis technologies to suit special needs. Articulatory synthesis technique has been shown to be one of the most promising solutions to produce synthetic speech. Since it is essentially based on the modeling of the human speech production mechanism, the articulatory control, therefore, provides an added degree of flexibility and controllability.

In this thesis, artificial neural network is used as the basic architecture of an articulatory type speech synthesizer for the production of Chinese language (specifically for Cantonese, which is a common dialect in Hong Kong). This method enables non-parametric speech templates to be stored implicitly as the synthesizer network parameters. Only a small number of speech templates are needed because individual phone is used as the basic units for synthesis. Furthermore, the non-linear properties of neural network also facilitate complex approximation for the relation between input control and the desired output. This neural network speech synthesis technique, thus, provides a bridge between the gap of articulatory synthesis and copy concatenation by adopting a non-parametric template concatenation approach with articulatory control.

It is well-known that vowel quadrilaterals have long been used in phonetics science for the analysis of speech sounds. In this work, the simplified articulatory space of the vowel quadrilaterals is employed as the basis of input control for the synthesizer networks. These articulatory parameters namely the **front-end position of tongue**

body, the **oral cavity openness** and the **lip-roundness** are believed to be sufficient for the description of most vowels.

Moreover, by making use of the knowledge in articulatory phonetics, **voice correspondence** is proposed to simplify consonant synthesis and to provide perceptually better phonetic segment (consonants or vowels) transitions for the synthesized output. Based on the place of articulation of target sounds, correspondences in vowel space are assigned to these sounds. The assigned correspondences are used as targets (interpolation control points) for the generation of articulatory parameter trajectories. Together with the neural network synthesis technique, improved transitions across phonetic segments and within compound vowels are made feasible in a phone-based synthesis approach.

In order to achieve better control of the timing properties during synthesis, statistical data are also collected to determine the inter-syllable pause duration in Cantonese phrases. Pause is one of the many different important prosodic features that governs the overall perceptual quality, especially naturalness, in spoken utterances.

As a result of the work in this thesis, the neural network approach is shown to be capable of synthesizing good quality speech signals. From the listening tests, the synthesized speech is found to have fairly high quality particularly in short segments where the importance of acoustic properties usually dominant. The articulatory control and voice correspondence have also been demonstrated to play a significant role in synthesizing speech that possesses a high degree of naturalness in transitions between different phonetic segments. Further improvement in the rhythm of the output speech is achievable through the employment of proper inter-syllable pause control.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The invention of writing and paper are certainly milestones in the development of human communication. History has also shown that mankind has successfully made wise use of the material around them or their own body parts to achieve effective (may be efficient) communication with people distant from them, both in space and time.

On the other hand, long before the history of writing, human has already been able to communicate in voice simply because there is a greater need for this. When the parties involved in the communication need instant information but they are busy with something else, voice will be a good choice. While the sender may not be able to post gesture or make symbols and the receiver may not be able to watch, their mouth and ears are usually free for communication. Therefore, speech communication is necessary in these situations for effective communication.

With a proper language system, speech communication has the advantage of being able to deliver much more information and even abstract idea as paralinguistic information. Posting gesture may not be sufficient for the delivery of much abstract idea and meaning. It may also be toilsome for communicating over a long period of time. With speech communication, these problems are alleviated. Moreover, the sender may pass paralinguistic information in parallel to the apparent linguistic information in the words. As the language system continues to develop, the effectiveness and efficiency of speech communication also improve.

During the past century, many technologies have been developed for efficient communication – speech coding, recognition and synthesis. The advent of these advanced

technologies for the distribution of information and communication has been made possible with speech. Although multimedia is now getting more popular for it can deliver information in various efficient and comfortable media, speech is still the choice in most cases. Its simplicity and speed make it more desirable for instant communication. Moreover, it has great versatility and higher ratio of perceivable information to data rate than most images. In most languages, its direct relation with text is also a desirable property. Last but not least, people simply have already got used to communicate in speech through their daily life and this makes it a natural choice for them.

1.1 Applications of Speech Synthesis

It is always a dream for the human beings to make robots that can imitate mankind in as many aspects as possible. This desire has driven much research in the area of speech synthesis for the goal of modeling the human speech production mechanism. Although, many works have been demonstrated, it is still far from ideal. After all, these limitations do not prevent people from using speech synthesis techniques in place where it is necessary and beneficial. In this section, several important applications will be introduced to show the wide applicability of this technology.

1.1.1 Human Machine Interface

Machine has improved the life of human by making it better and easier. However, human interfacing is always a major problem. Buttons, switches, handles, lamps, bells, displays, and the likes have been used for a very long time as simple interface devices. However, as the complexity of the system increases, the control, output and feedback to users also become more complicated. Simple interfaces are becoming inefficient and insufficient. There is a growing need for more natural and efficient way of machine interface for practical purpose.

In human machine interface, speech is one of the most efficient channels when used properly. Simple systems that interfaces with the general users may have their speech interfaces. Some public addressing systems like announcement in elevators, trains and public area as well as those simple sound activated switches are good examples. However,

without an "intelligent brain", their simplicity can only provide interfaces of limited functions.

In order to communicate with human in speech, machine must be able to recognize and understand the speech of the talker. There has been much work in this area and a high recognition accuracy can be achieved under controlled conditions which makes limited functions speech input already feasible.

On the other hand, for people to receive information back from the machine in voice, speech synthesis is necessary. This is especially important in cases where users are eye-busy with some other information. Furthermore, sound is an effective way of drawing people's attention. As time evolves, more complex response from the machine will make current technology insufficient. Advanced techniques must be developed to meet this growing demand. When the complexity of machine response gets higher, it is desirable to constitute idea and present it as speech signal through the control of some speech production mechanism. These applications place much room for further development in the area of synthesis technique for human machine interface.

1.1.2 Speech Aids

Certainly, technologies in speech synthesis have given people many convenient and natural interfaces. Needless to say, there are also applications where speech synthesis is not just for convenience, but it is a must. In applications like speech aids for people with speech impairment, speech prosthesis has brought light to them.

Speech prosthesis has long been investigated and many such systems have already been demonstrated. However, as a particular application of speech synthesis, it also faces some of the common problems that most speech synthesis applications have. From early time systems that perform simple playback of recorded sentences and words, to parametric speech models, the controllability has been improved a lot. Unfortunately, the system complexity or naturalness of the output speech are sacrificed in some sense. In order to give a better systems for speech prosthesis, properties like controllability, system complexity and output naturalness must also be taken into account.

Due to the limitation of technology, nearly all speech prosthesis systems available are still text to speech systems. The patients are actually talking through writing and

reading. However, as an ideal prosthesis system, it should synthesize speech signal by following instructions from speakers' intention and idea. Biological signals such as electromyograph (EMG) signal has been used for the generation of articulator controls and then to speech sound [1]. Indeed, techniques such as articulatory synthesis have been regarded as the promising solution. Ultimately, patients should be able to talk with their minds as general people do.

1.1.3 Text-To-Speech (TTS) system

Among the various applications of speech synthesis, text to speech system is perhaps one of the most attractive areas. It is because of the potential market and the vast practical use. TTS performs the task of changing text data to speech signal. It is useful for speech presentation of information that are used to be delivered as text. TTS can be used for reading out text data either for proof-reading purpose or convenient parallel interface to users.

More importantly, it can be employed as value-added service to those existing information distribution services. It enables information to be transferred over voice lines as speech. It provides fast to market value-added information service. Users can now gain easy access to information that are used to be accessible only with modems and display terminals or with operator interfacing. This gives great convenience to the users and reduces labour cost for the service providers.

In addition, TTS also gives flexibility to the information provider because data are kept in text form which allows easy and prompt changes, updates and corrections when compared to those pre-recorded speech information distribution service.

1.1.4 Speech Dialogue System

Ideally, we would prefer communication with machine using free speech. The machine will understand speech input, take appropriate action and feedback in spoken language. This understanding of speech, constituting of idea and generation of speech signal comprise an ideal dialogue system. These kind of dialogue systems are ideal for the human machine interfaces.

A typical dialogue system is the machine translation system. Speech spoken in one language is understood. The idea and meaning is rephrased in another language and output as speech signal. Moreover, information distribution service can now have information stored as knowledge database. Machine will deliver the content as the targeted language in response to spoken requests or queries. The capability of using knowledge database information also allows better use of the information inventory. However, it presents a more difficult task for speech synthesis.

In all these kinds of dialogue systems, the requirement on the controllability of the speech synthesizer is the most stringent. Since the speech output is constituted from idea as in a human brain, similar degree of control as the human speech production system should be allowed for generation of the speech. Synthesis algorithms that lack controls may be found inappropriate in these applications. Therefore, there is an increasing demand for easily controllable synthesis techniques. In future, with the success of this kind of synthesis system, machine may make up speech from idea and deliver the information to user. We will then be getting closer to the ideal of imitation of human speech production mechanism.

1.2 Current Status in Speech Synthesis

Nowdays , there are many different techniques for speech synthesis. Depending on the specific area of application, different techniques may be used. Among those common ones, several main stream techniques are described below.

1.2.1 Concatenation Based

The simplest speech synthesis method is based on template concatenation. Recorded speech signals are stored for later playback upon request. There are also time domain synthesis techniques such as that proposed by Costello and Mozer [2]. It attempts to concatenate speech templates that are modified to zero phase segments.

The main drawbacks of concatenation methods are the large storage requirement and limited output flexibility. While they have the advantage of retaining much of the naturalness through the recorded speech templates.

By making the templates smaller, the storage requirement is relaxed and the flexibility in output is increased. More output variations are made possible through the use of shorter templates in concatenation. However, the perceptual quality is lost due to the discontinuity at transitions between templates and the lack of context-dependent variations in acoustic properties of templates.

Greater variations in templates can be achieved by using more templates with different acoustic properties. However, the storage will then grow larger and larger. Although vast storage is not a key problem in most large systems, adoption of this approach will hinder implementation in small systems. Moreover, the large amount of templates also brings about problems in invertory searching which is usually time-consuming. In order to select proper template size, there are systems being developed for automatic template size selection from template inventory [3]. In addition, the advent of pitch-synchronous overlap add (PSOLA) [4] technique also allows simple and efficient change of the pitch profile and timing of time domain speech signal.

Generally speaking, this type of synthesizer is popular for its simplicity and commonly used in fixed systems such as announcement and simple query systems.

1.2.2 Parametric Based

The most popular synthesis technique nowadays is the parametric synthesis technique such as formant synthesis and linear prediction coding (LPC) based methods. They provide parametric system model for speech synthesis. For formant synthesis, the amount of templates stored can be reduced by providing rules for the variations and trajectories of the formants and bandwidths. Using formants and bandwidths as templates is found to be effective for the reproduction of speech signal. However, the spectral fine details of the speech signals may be lost. These approaches usually find application in many different areas with moderate requirement on the acoustic properties. In general, naturalness of speech output from formant synthesizers are improved through the incorporation of controlling rules for prosodic properties such as pitch and timing. The DECTalk commercial synthesis system is a typical example that is based on the KlattTalk proposed by Dennis H. Klatt. Many more similar systems have also been put to market recently.

1.2.3 Articulatory Based

Another synthesis technique that is popular in research area rather than real world application is the articulatory synthesis method. It has not yet been widely used in commercial products mainly due to its complexity. This method tries to simulate the actual vocal tract and voice source reliably. Controlling this type of synthesizers requires specification of the articulators positions or vocal tract area function that changes the vocal tract properties. For better control of these synthesizers, people are working on the derivation of controlling area function from speech signal – the inverse problem. Even though this synthesis technique is complicated, it still attracts much attention because it has great potential to be the ultimate solution for the speech synthesis problem. With a reliable model for the vocal tract and glottal excitation source, good synthesis result is possible. This physically related control also provides much flexibility and controllability.

Recently, there are also many works trying to incorporate articulatory control to other synthesis techniques such as copy concatenation [5] or formant synthesis [6]. Advantages of different techniques are being integrated into single methods.

1.2.4 Application of Neural Network in Speech Synthesis

Much effort has been spent to apply ANN to facilitate the production of high quality speech signal. The advantages of using neural network are that knowledge can be incorporated through the training stage.

Moreover, certain degree of non-linearity can be incorporated such that more complicated models can be built. The neural network approach also has the potential of parallel implementation that renders real time application.

Different issues relating to the application of neural network to speech synthesis has been studied extensively and are summarized as follows.

Text-To-Transcription Conversion Neural networks are used for conversion of orthographical text to phonetic transcription. NetTalk [7, 8, 9, 10, 11, 12] and NetSpeak are famous examples. This is a crucial problem in text to speech conversion for achieving reliable machine reading of orthographic text. The complicated relationship between text and pronunciation is modeled by the neural network.

Acoustic Synthesis They are used for conversion of phonetic transcription to digital filter parameters for synthesis purpose [13, 14, 15, 16, 17, 18, 19, 20], particularly in the generation of formant synthesizer parameters and linear prediction filter parameters such as reflection coefficients (PARCOR), line spectrum pairs (LSP).

Inverse Problem Neural networks have been used to enable back conversion of speech signal to the vocal tract area function [21, 22, 23, 24]. These results are useful for both speech science as well as speech synthesis. They are especially useful for those synthesis techniques that employ articulatory control.

Vocal Tract Model Use of neural networks to model the human vocal tract have been investigated [25, 26]. In addition, they are also applied for the generation of articulatory motion from the phonemes [27] or from EMG signal [1] for synthesis purpose.

Other Areas Other applications such voice conversion [28, 29], learning of speech waveform dynamic changes [30] and spectral interpolation by phonemes vectors [31] have also been demonstrated.

1.3 The Proposed Neural Network Speech Synthesis

In this thesis, artificial neural network is proposed to synthesize speech signal through training with spectral templates of phone units. The synthesizer networks are controlled by articulatory control parameters. It provides a novel approach for the phone-based concatenation synthesis with articulatory control.

1.3.1 Motivation

In small unit template-based concatenation synthesis system, the major problems are templates transitions and difficulty in modeling spectral variations from the stored templates. Discontinuities are common in the junction of concatenated speech templates. In addition, same segment of speech may have many different variations under different context and conditions. If a small amount of templates are used, these variations cannot be represented vividly. This results in synthetic output that lacks proper variations.

In order to overcome these problems, larger amount of speech templates are usually employed for modeling the speech variations and different types of transitions. This, however, brings about other problems such as storage and searching for proper template units from inventory.

Articulatory synthesis, on the other hand, is believed to be of great potential for the speech synthesis applications. It provides much flexibility and controllability for the synthesis process. It is effectively a simulation of real speech production process and therefore reliable modeling will guarantee good quality output. For this reason, it is used as the synthesis control in some speech synthesis techniques [5, 98].

1.3.2 Objectives

In this thesis, artificial neural network is proposed for the synthesis of speech. The goals of this approach are as follows,

Small Template Concatenation Phone-based approach for concatenation synthesis is adopted to achieve a synthesis method with better flexibility and only small storage requirement for templates.

Reliable Spectral Template Mapping Reliable reproduction of the spectral speech templates is achieved by the mapping of the synthesizer network through the training process.

Flexible Synthesis Control Flexible synthesis control is often desirable. It is achieved by providing simplified articulatory parameters to control the synthesizer network for the generation of speech signal.

Template Transition Approximation Transitions for the phone templates are approximated by the synthesizer neural network output. The synthesizer network performs non-linear approximation for the transition regions between trained templates. This also provides control model for acoustic properties variations of speech templates in output utterance.

In summary, artificial neural network is used in an attempt for the mapping of the articulatory control parameters to the non-parametric spectral information. It tries to make use of the non-linearity of the network to learn the relationship between the control parameters and the speech spectral information that is provided through the training process. While intuitive and flexible control is achieved by the use of simplified articulatory parameters as synthesis control.

1.4 Thesis outline

This thesis is divided into three main parts, prologue, core and the epilogue. Prologue includes the introduction and some background knowledge. The applications of the technology and the problems handled in this work will be introduced. It provides the reader with a general idea of the goal of this work. Chapters are also devoted to the introduction of basics on the speech science and synthesis technology. Brief overview of the related knowledge in linguistics and speech synthesis is given.

The core will elaborate the details of the work. These include discussion on the application of artificial neural network for speech synthesis, the implementation of prototype synthesizer networks and description on the proposed articulatory control incorporated in the synthesizer network. In addition, statistical result of the inter-syllable pause duration of Cantonese phrases is also included. It provides the basic guideline for inter-syllable pause duration control, and it is shown to be one of the important prosodic features in natural speech utterance.

Finally, the epilogue will draw conclusion on the work and also propose some direction for further work on this particular speech synthesis problem.

Chapter 2

Linguistic Basics for Speech Synthesis

Linguistics, a science of systematic study of human languages. It is made up of many branches employing different approaches to deal with various problems of the science. This field of knowledge gives much insights and applicable information to the area of speech processing. In particular, with the application of linguistic knowledge and tools, the research and development of speech synthesis technology has been greatly improved. In order to facilitate better speech synthesis with more flexibility, multi-disciplinary work is necessary.

2.1 Relations between Linguistic and Speech Synthesis

The general relationship between linguistics and speech synthesis is summarized in figure 2.1. In a complete speech synthesis system, idea is first formed or given. The task of the system is then to present these information in the form of speech signals to the user. Based on the semantic meaning of the words in a lexicon, words are first searched according to the content of the outstanding information. Furthermore, the system may also apply pragmatic rules based on the context of the information and make selection among the possible list of choices. Subsequently, proper word-form is chosen for the selected words according to the word positions and the syntactic rules of the targeted language [1]

After the word selection process, these words are converted to acoustic signals which is a pertinent process in synthesis system. The actual pronunciation of the words are

[1]The order of the processes is not crucial and may vary in different implementation.

formulated according to phonological rules and data in the transcription database. The speech signal is then generated according to the phonetic properties represented by the transcription.

After all, the architecture shown is an illustration of an ideal synthesis system. In many practical implementations, only several important components are incorporated for specific requirement and applications .

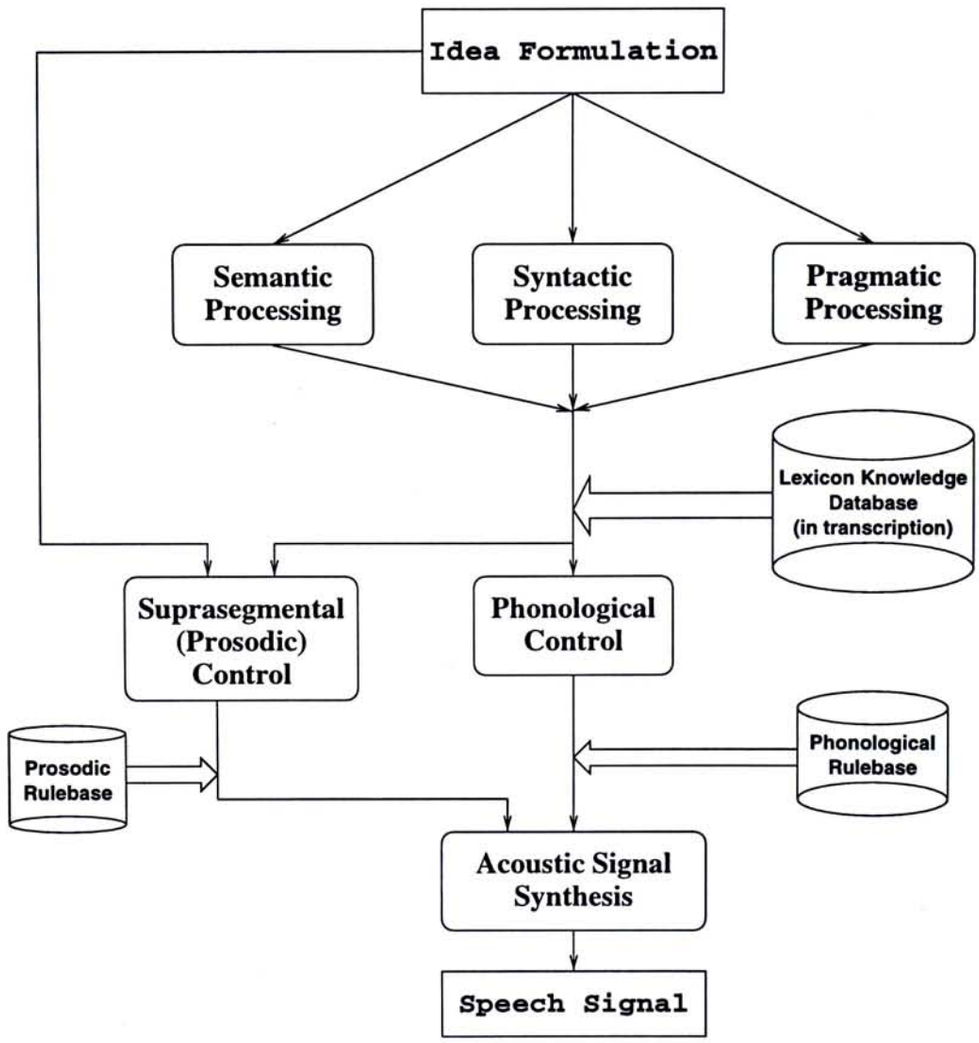

Figure 2.1: The diagram shows the relationship between linguistics and speech synthesis. The real human process of speech formulation runs from top to the bottom. In speech synthesis, progress is proceeding in a bottom-up direction with increasing difficulty.

2.2 Basic Phonology and Phonetics

2.2.1 Phonology

Phonology describes the process how speakers formulate speech utterances. It also describes the system of units for the formulation. It is the *study of the sound system of a given language.* It is also a study of the analysis and classification of phonemes – the minimal set of phonological symbols.

From the linguistics point of view, language is formed hierarchically. Sentence is made up of phrases and the phrases are from words. Each word consists of one or more syllables while syllables are comprised of phonemes. Phonemes are realized as speech signal in form of phones that represent the phonemes [Figure 2.2]. From the word level

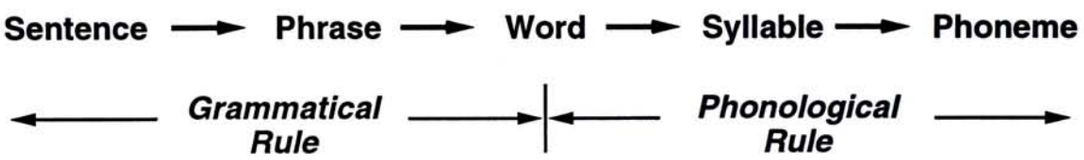

Figure 2.2: The hierarchy of units made up a sentence. Units are getting smaller and smaller until phonemes which is realized acoustically as phones.

and upwards, the grammatic rules determine the formation of the sentence. Those levels from words and below are determined by phonological and phonetic rules. Some of these may be language oriented while some are physically constraints.

Phoneme (音位) refers to the elements of a minimal set of symbols that are capable of unambiguous description of word pronunciation in a language. They are considered to be the minimal phonological separable units. Each phoneme may represent one particular sound or a group of similar sounds. The ambiguity is usually resolved by the phonological characteristics of the particular language.

Syllable (音節) lies between the phoneme and word level of segmentation. It consists of **onset** and **rhyme** [Figure 2.3]. The rhyme may further be separated into **nucleus** (or **center**) and the **coda**. Nucleus is the essential part of a syllable while the onset and coda may or may not exist. For some tonal languages, such as Cantonese, each syllable is further characterized by an additional features, the lexical tone. Lexical tone

is realized in form of pitch, amplitude and duration of the syllable. Therefore, the same set of phonemes may be referring to different meaning for different lexical tones.

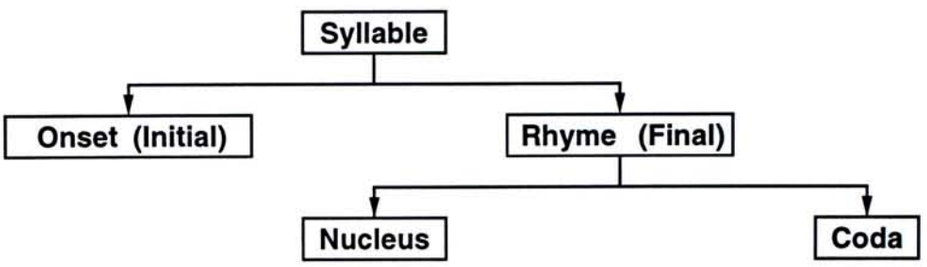

Figure 2.3: Units that make up a syllable

2.2.2 Phonetics

Phonetics is the study of pronunciation, the sound actually produced. It is a view of speech sounds considered in isolation from any languages.

From the phonetics point of view, the segmentable parts of the spoken syllables are known as segments. There are two main categories of segments, the **consonants (C, 輔音)** and the **vowel (V, 元音)**. Consonants are often characterized by the noise like properties (*/s/, /f/, /b/* etc.), while vowels are voiced segments with quasi-periodic acoustic properties (*/a/, /e/, /u/, /i/, /o/* etc.). Moreover, they are usually the energy carriers of syllables. Nuclei in syllables are usually made up of vowel segments. Depending on the language, the structure and number of constituting segments in a syllable may vary (V, VC, CV, CVC, CVC, etc). In real speech signals, phonemes are physically realized as **phones (音素)**. Phone is the smallest physically separable unit in the speech. For a particular phoneme, its variations in realization are known as **allophones (音位變體)**.

In contrast to simple vowels (簡單元音), there are also compound vowels (復合元音). Simple vowels maintain the spectral properties in a quasi-stationary states over the whole segments, while compound vowels are made up of more than one simple vowels as components. Within a compound vowel, the spectral properties change from that of one vowel to others. **Diphthongs (雙元音)** are compound vowels that made up of two simple vowel components and **triphthongs (三元音)** are made up from three.

2.2.3 Prosody

In general, properties regarding the segmentable units are referred to as segmental features. In a speech utterance, those features that are spread over segments are known as **suprasegmental features** or **prosodic features**. Some important prosodic features are pitch variations, timing, amplitude profile. In spoken utterances, prosodic features may provide additional paralinguistic information in parallel to the apparent information from wordings.

In a sentence, **intonation** is the pitch properties that spread over the sentence and slightly modify the pitch of syllables. **Stress** determines the position in a sentence where attention is drawn by exaggerating amplitude and a rising pitch. Timing is another important feature in speech. It refers to the duration of segments, inter-segment transition as well as the pause between the phonetic segments. Timing properties of speech segments usually determine the perceived rhythm and naturalness of uttered speech.

2.3 Transcription Systems

Transcription is an important topic in linguistics. In most languages, there is not any easy and direct relation between word pronunciation and writing. Also, there are occasions that unprecedented or unwritten sounds are required to be described. For these purposes, methods for the transcription of pronunciation are needed and these lead to the advent of phonetic transcription. There are many different systems of transcribing sounds all over the world. Based on Daniel Jones's transcription method (and the cardinal vowels diagram), the International Phonetic Association [32, 33] has made up the International Phonetic Alphabets, IPA, which then becomes the most widely adopted and based system. It is made up from romantic characters, some Greek characters as well as additional rotated or invented characters.

There are some basic principles [32, 33] that make a transcription system more practical and universal. Firstly, for a particular language, *different sounds for distinguishing different words should use different alphabet representation without diacritical marks* [2]. Secondly, *similar sounds not used for distinguishing words may use the same character representation* and diacritical marks are used to distinguish them when needed.

The IPA transcription scheme has the narrow form and the loose form for different applications.

Loose form – It is **phonemic transcription**. It emphasizes on phonological contrast of pronunciation. It **unambiguously** describes the pronunciation for the target language using a **minimal** and unambiguous set of letters [32, 34]. Phonemic transcriptions are usually denoted by enclosing with the pair of "/ /".

Narrow form – It is used for detail description of the pronunciation. It aims at a more elaborative and in depth description of the language. It can be used for studying dialects and comparing languages. Diacritic marks are used in additional to the basic set of phonetic alphabets to modify the basic sound represented by the alphabets (e.g. nasalization, tones etc.). It is also known as phonetic transcription. In writing and printing, phonetic transcriptions are enclosed with "[ ]".

[2] Diacritical marks are additional symbols to the phonetic alphabets for describing small changes in their pronunciation. Their use of is limited to cases : i) denoting length, stress and pitch; ii) representing minute shades of sounds; iii) its introduction will obviates the need of additional alphabet.

2.3.1 The Employed Transcription System

The IPA system of transcription has the advantage that it is popularly used. However, some of the letters used are not ASCII characters making it difficult to be typed on conventional keyboard. Many modified transcription schemes have been developed to overcome this problem, such as the DARPABET (ARPABET) for American English. For Cantonese, there has also been some schemes developed with similar goal. In The Speech Processing Group of Department of Electronic Engineering, the Chinese University of Hong Kong, a modified version is adopted for Cantonese transcription. The major differences in the transcription scheme include

- Restricting the initial and final part of the transcription to be two alphabets long.
- Allowing an additional digit for representing the lexical tone of the syllable.
- All transcription symbols are ASCII and case insensitive, thus making labeling, transcription and documentation easy.
- Each syllable is either 4 (without tone) or 5 (with tone) character long. Therefore, it is ideal for naming files on PC running MS-DOS [3] operating system. The remaining 7 or 6 letters can be for other purpose such as speaker and trial identification.

This is basically a phonemic transcription scheme. Unless otherwise specified, this scheme will be assumed throughout this thesis.

For comparison, some popular Cantonese transcription schemes in Hong Kong are tabulated in parallel with the closest IPA symbols in table 2.1. These include the IPA, Yale scheme, Sydney Lau's scheme, scheme of the Linguistics Society of Hong Kong (LSHK) 1993 and the adopted CUHK-EE scheme.

[3] MS-DOS is a trademark of MicroSoft Corporation

IPA	Yale	Sydney Lau	LSHK	CUHK-EE	Remarks
Initial					
				nu	Null initial
p	b	b	b	b_	
t	d	d	d	d_	
k	g	g	g	g-	
k^w	gw	gw	gw	gw	
p^h	p	p	p	p-	
t^h	t	t	t	t_	
k^h	k	k	k	k_	
k^{wh}	kw	kw	kw	kw	
s, ʃ	s	s	s	s_	/ʃ/ before /y/
f	f	f	g	f_	
h	h	h	h	h_	
ts^h, $tʃ^h$	ch	ch	c	ts	/tʃ/ before /y/
ts, tʃ	j	j	z	dz	/dʃ/ before /y/
l	l	l	l	l_	
m	m	m	m	m_	
n	n	n	n	n_	
ŋ	ng	ng	ng	ng	
w	w	w	w	w_	
j	y	y	j	j-	
Vowel/Nucleus					
i, ɪ	i	i	i	i_	/ɪ/ in velar (/ɪng/,/ɪk/)
e, ɛ	e	e	e	e_	/e/ in diphthong /ei/
a	aa	a	aa	a_	
y	yu	ue	yu	y_	
œ ɸ	eu	eu, euh u	oe eo	j-	in simple vowel, velar (/œng/,/œk/) in diphthong /ɸy/, alveolar (/ɸn/,/ɸt/)
ɔ, o	o	o, oh	o	c_	/o/ in diphthong /ou/
ʊ u	u	u oo	u	u_	in velar (/ʊŋ/,/ʊk/) in simple vowel, alveolar
ɐ	a	a	a	r_	
Coda					
p	p	p	p	p	
t	t	t	t	t	
k	k	k	k	k	
m	m	m	m	m	
n	n	n	n	n	
ŋ	ng	ng	ng	5	5 is used because it is pronounced as /ng/ in Cantonese

Table 2.1: The symbols in some selected Cantonese transcription schemes are tabulated together with the related IPA symbols.

2.4 Cantonese Phonology

Cantonese is a monosyllabic tone language. It is a Chinese dialect that is used in some of the South China areas and overseas. It is also the most commonly used language in Hong Kong. From the linguistics' point of view, it is important in the sense that it preserves some ancient Chinese sounds and words in its vocabulary. We do not attempt to give a thorough review because it is really beyond the scope of this work. Interested reader are referred to the following references [35, 36, 37]. On the contrary, some fundamentals about the phonology of Cantonese will be introduced such that basic concept can be developed to make it easier to go on to the later chapters.

2.4.1 Some Properties of Cantonese

In Cantonese, there are many different characteristics that are unique to this Chinese dialect. It is one of the dialects that preserve the nasal sounds and used extensively in its syllables. The nasal consonants, /m/, /n/ and /ŋ/ may appear as syllable initial or final coda. Phonologically, a single Cantonese syllable is also allowed to possess both nasal initial and coda. In addition, there are also syllabic /m/ and /ŋ/ where these nasals are syllables themselves. Alveolar, labial and velar are important articulation position in Cantonese consonants. These articulation positions are characteristics to the nasal consonants and stop codas which in turn form important parts in the dialects.

Place of articulation	**Nasal consonants**	**Stop codas**	Nearby vowel
Alveolar	/n/	/t/	/i/
Labial	/m/	/p/	/u/
Velar	/ŋ/	/k/	/ɔ/

Since Cantonese is monosyllabic, speech in Cantonese is made up from string of monosyllables. Phonologically, besides the onset-rhyme segmentation, we usually refer the building blocks of most Chinese languages as **initials** (**聲母**) and **finals** (**韻母**). The full list of Cantonese initials and finals is given in the Appendix A.1.

2.4.2 Initial

Initial is the first part of a syllable. In Cantonese, there are totally 19 initials plus the null initial – syllables without initial. The variety of initials includes fricative – /f/, /s/, /h/; affricative – /ts/, /dz/; unaspirated plosive – /b/, /d/, /g/, /gw/; aspirated plosive – /p/, /t/, /k/, /kw/; nasal – /m/, /n/, /ng/; lateral liquid – /l/; and semi-vowel glides (the approximants) – /w/, /j/. They are tabulated in table 2.2 [38]. [4]

	Bilabial	**Labial dental**	**Alveolar**	**Palatal**	**Velar**	**Glottal**
Plosive	/p/, /b/		/t/, /d/		/k/, /g/ /kw/, /gw/	
Nasal	/m/		/n/		/ng/	
Fricative		/f/	/s/			/h/
Approximant	/w/			/j/		
Lateral			/l/			
Affricatives			/ts/, /dz/			

Table 2.2: Table of Cantonese initials and the manner and place of articulation

Among these initials, they have some characteristic properties. Some of these properties are listed as follows.

- /b/, /d/, /g/ are unvoiced unaspirated plosive consonants. They are the unvoiced versions of the /b/, /d/, /g/ in English. Their IPA symbols are /p/, /t/, /k/. Similarly, /p/, /t/, /k/ are unvoiced aspirated plosive, and their IPA symbols are /p^h/, /t^h/, /k^h/ respectively. In Cantonese, initials /b/, /p/ does not form syllables with lip-rounded vowels. In addition, initial /b/ is only followed by /ɪ/ and not /i/ [5].

- /m/, /n/, /ng/ have similar articulatory position as /b/, /d/, /g/. and their IPA symbols are /m/, /n/, /ŋ/ respectively. In Cantonese, the labial /m/ does not form syllables with lip-rounded vowels.

[4] Some people use 22 initials plus the null initial. The /s/, /ts/, /dz/ before lip-rounded vowels and others are differentiated.

[5] There are new Cantonese sounds that are created from English such as "BB", /bi/ /bi/ (means "baby") and /bʊm/ (means "pump").

- /f/ is the unvoiced labial-dental fricative consonant with IPA /f/.
- /l/ is the unvoiced alveolar lateral liquid with IPA /l/.
- /h/ is a vocal aspirated fricative consonant with IPA symbol /h/. Syllables with initial /h/ are usually associated with negative voice-onset-time and the initial /h/ may sometimes be made voice-like by the intra-syllable coarticulation from its finals or inter-syllable coarticulation from preceding final.
- /ts/, /dz/ are affricative and /s/ is fricative consonant. They have similar phonological properties in Cantonese that each of them groups two different initial phones. /ts/ is the unvoiced aspirated affricative of IPA /ts^h/ and /ʧ^h/. /dz/ is the unvoiced unaspirated affricative of IPA /ts/ and /ʧ/. And /s/ is the unvoiced alveolar fricative of IPA /s/ and /ʃ/. They can group two different phones together is due to the characteristics of Cantonese that /ʧ^h/, /ʧ/ and /ʃ/ occur exclusively before lip-rounded vowels and not for /ts^h/, /ts/ and /s/.
- Semi-vowels glides /w/ and /j/ correspond to the IPA symbols of /w/ and /j/. They usually have no friction or just a little bit friction.
- /gw/ and /kw/ are the lip-rounded version of the velar plosive /k/ and /k^h/. They have the IPA symbols of /k^w/ and /k^{wh}/. Phonetically, it is similar to a velar consonant immediately followed by the glide /w/ in pronunciation.
- Null initial means syllables that have finals only. In practice, there is a trend in Cantonese that they are velarized. It either possesses some minor velar starting phenomenon or even becomes "velar (/ng/) like" [35, 39].

2.4.3 Final

In Cantonese, there are 53 different finals. Among these 53 finals, they come from four categories : *simple vowel*, *diphthong*, *vowel nucleus with coda* and *nasal*. A brief description of each category is given below and the corresponding lists are given for reference.

- **Simple Vowel**

  In Cantonese, there are eight different vowel phonemes. However only seven of them are used as simple vowels in finals. The seven simple vowel finals, namely /a/, /e/, /i/, /c/, /u/, /y/ and /j/, cover most of the vowel phonemes in Cantonese while the phoneme /r/ occurs in diphthongs only.

- **Vowel Nucleus with Coda**

  In those finals making up from vowel nuclei and codas, the codas are either nasal or stop consonants. Possible nasal consonants are /m/, /n/ and /5/. Stop consonants are the unvoiced unaspirated /p/, /t/ and /k/ where all of them are always associated with the entering tone. Valid finals in this category are tabulated in table 2.3 [40].

	Labial		**Alveolar**		**Velar**	
	Nasal	**Stop**	**Nasal**	**Stop**	**Nasal**	**Stop**
	m	p	n	t	ŋ	k
i ɪ	im	ip	in	it	 ɪŋ	 ɪk
y			yn	yt		
u ʊ			un	ut	 ʊŋ	 ʊk
ɛ					ɛŋ	ɛk
ɸ œ			ɸn	ɸt	 œŋ	 œk
ɔ			ɔn	ɔt	ɔŋ	ɔk
ɐ	ɐm	ɐp	ɐn	ɐt	ɐŋ	ɐk
a	am	ap	an	at	aŋ	ak

Table 2.3: Cantonese finals making up from vowel-nucleus with coda is tabulated. (IPA symbols is used here to illustrate some of the phonological properties in Cantonese.)

- **Diphthong**

  In Cantonese, there are two types of diphthongs, the labial terminated and the alveolar terminated. The labial terminated diphthongs are ended with the close-back vowel /u/ and the alveolar terminated diphthongs are ended with the close-front vowels /i/ or /y/. They make up a total of ten different diphthongs for the dialect and they are tabulated in table 2.4.

	starting vowel						
	/a/	/e/	/i/	/c/	/u/	/j/	/r/
Close-and-Front /i/ or /y/	/ai/	/ei/		/ci/ (/ɔi/)	/ui/	/jy/	/ri/
Close-and-Back /u/	/au/		/iu/	/cu/ (/ou/)			/ru/

Table 2.4: The Cantonese diphthongs table showing all Cantonese diphthongs with the corresponding pairs of constituting vowels. (The bracketed ones are IPA for showing the slight variation of /c/ in different diphthongs)

- **Nasal**

  In Cantonese, there are two syllabic nasal sounds that are used as finals. They are /m/ and /ŋ/. However, these finals do not occur very often in the Cantonese syllabary and is included here for completeness.

Phonological rules in final formation

In the formation of finals from the vowels and/or codas, there are some phonological limitations. The selection of phoneme symbol is based on some of these unique phonological properties. Some of these constraints are listed as follows.

- For the lip-rounded vowels /u/, /ɔ/, /j/ and /y/, they never form finals with the bilabial codas /p/ and /m/ (except informal colloquial phrases);
- Finals with velar codas, /k/ and /ŋ/ are usually made from shorter vowel nuclei compared to those for alveolar /t/ and /n/;
- /y/ only forms finals with alveolar /n/ and /t/;

- /a/ and /r/ are the two only vowel nuclei that can form finals with all codas.
- /ɔ/ does not form diphthongs with /u/ while /o/ only forms with one with /u/ [A single phoneme symbol, /c/ is used];
- /ɸ/ forms finals with alveolar (/n/, /t/) while /œ/ form finals with velar (/ŋ/, /k/) codas only [A single phoneme symbol, /j/ is used];
- /i/ forms finals with alveolar (/n/, /t/) and labial (/m/, /p/) while /ɨ/ form finals with velar (/ŋ/,/k/) codas [A single phoneme symbol /i/ is used];
- /u/ forms finals with alveolar (/n/, /t/) and /ʊ/ with velar (/ŋ/, /k/) codas [A single phoneme symbol /u/ is used];
- /ɛ/ nucleus associate with velar (/ŋ/, /k/) codas which is also mutually exclusive to /e/ that form diphthong with /i/ [A single phoneme symbol, /e/ is used]. In colloquial phrases, /ɛ/ also forms finals with labial (/p/) coda.

2.4.4 Lexical Tone

Cantonese is a tonal language. It has a total of nine different lexical tones （九聲）. Tones in Cantonese play important role in that they are necessary for recognizing the syllable under consideration. Although the syllables themselves provide much clue for recognition from the contextual and acoustic properties, correct recognition of tones help reducing much of the ambiguity and uncertainty.

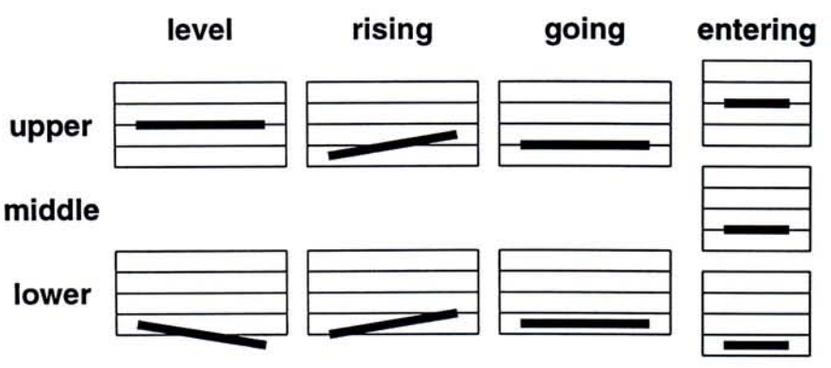

Figure 2.4: The characteristic pitch levels and profiles of the nine tones in Cantonese in the musical scale.[41]

		Tone Marks		S.L.Wong[41]	廣拼[35] [a]
upper-level	(陰平)	˥x	55 (53) [b]	1	1
upper-rising	(陰上)	˧˥x	35	3	2
upper-going	(陰去)	˧x	33	5	3
lower-level	(陽平)	˩x	11	2	4
lower-rising	(陽上)	˩˧x	13	4	5
lower-going	(陽去)	˨x	22	6	6
upper-entering	(陰入)	˥x	5	7	7
middle-entering	(中入)	˧x	3	8	8
lower-entering	(陽入)	˨x	2	9	9

Table 2.5: Some of the Cantonese tones labeling schemes are listed. The scheme of 廣拼 is used in this thesis.

[a] Cantonese Transcription Scheme (廣州話拼音方案)

[b] Tone 1 includes both level (55) and level-drop (53) pitch profile. The use of which profile depends on particular syllables and speaker habit. Reader may refer to [35] for further information.

The lexical tones are realized by their characteristic pitch level, pitch profile, intensity and timing. The tones are categorized into two main categories, upper and lower (陰、陽). For each of these categories, it is subdivided into level (平), rising (上), going (去) and entering (入) tone. With the additional middle entering tone, it makes up a total of nine different tones. The approximated pitch levels and profiles are sketched in figure 2.4.

The level and going tones in general have level or level-drop pitch profiles. The rising tones have the characteristic rising profile and entering tones are characterized by the short durations. In addition, all syllables in entering tones are associated with special codas – /p/, /t/, /k/. The upper-entering, middle-entering and lower-entering tones have the similar pitch level as the upper-level, upper-going and lower-going tones respectively. Some popular schemes for lexical tones transcription are list in table 2.5.

2.4.5 Variations

In continuous utterance, there are many variations in the uttered speech signals due to different reasons. These variations are especially common in informal and colloquial conversation.

Tone Variation

The tone of the syllables in an utterance may be changed under various conditions. In general, there are two main causes [35]:

1. **Coarticulatory**

   The change in tone is caused by the tone of adjacent syllables.

   - Tone 1(53) + Tone 1(55) change to Tone 1(55) + Tone 1(55), e.g. /h_j51/ /dziu1/, 香蕉 (means "banana");
   - Tone 1(53) + Tone 1(53) change to Tone 1(55) + Tone 1(53), e.g. /g_c51/ /s_an1/, 江山 (means "landscape");
   - Tone 1(53) + Tone 7(55) change to Tone 1(55) + Tone 7(55) e.g. /s_an1/ /g_uk7/, 山谷 (means "valley");
   - For some phrases of relative name or children colloquial phrases, the tones of first syllable are changed to Tone 4(11) regardless of the original tones, e.g. /d_ri4 d_ri2/, 弟弟 (means "brother").

2. **Habitual**

   The change of tone is habitual in particular society and especially common in informal and colloquial conversation.

   - Tone is changed to Tone 1(55), e.g. /h_c_4 l_an1/, 荷蘭 (means "Holand");
   - Tone is changed to Tone 2(35), e.g. /t_c_1 h_ai2/, 拖鞋 (means "slipper").

Acoustic Variation

The acoustic variations refer to those changes due to the inter-syllable coarticulatory effect. In general they can be classified into two main types :

1. **Assimilation**

   Due to the articulatory positions of the initials and finals in phrases. The initial of next syllable may be assimilated by the final of previous syllable and vice versa. The effect is more prominent in fluent continuous utterances. For example, in /b_ak8/

/dzuk7/,百足(means "centipedes"), the final coda velar /k/ may be assimilated to alveolar /t/ and becomes nearly /b_at8/ /dzuk7/.

2. **Merging**

   In fast continuous utterances, especially colloquially, two syllables may be merged into one. The merged syllable may possesses unusual initial. For example, the colloquial phrase, /h_rm6/ /b_a56/ /l_a56/ (means "all") may be merged to /h_rm6/ /bla56/ with merged initial /bl/.

2.5 The Vowel Quadrilaterals

In order to set up reference for the study and identification of sounds in languages, British phonetician Daniel Jones (1881-1967) devised some standard reference points in the articulatory space as standard vowels. Their assignments are based on both *articulatory* and *auditory* judgment. These reference vowels are also known as **cardinal vowels**. They are used as reference for phonetic study rather than to represent any real vowels. With these cardinal vowels, phoneticians can investigate and compare the utterance of different languages through the deviations of the sounds from those cardinal vowels.

There are two set of cardinal vowels, the primary and secondary series of cardinal vowels. The primary cardinal vowels correspond to the sounds uttered without lip-rounding. While the secondary series correspond to those with lip-rounding. These vowels are usually displayed in the form of the **vowel diagram (vowel quadrilateral)**. The coordinate system of the vowel quadrilateral is made up of the front-back position of the tongue body in horizontal direction and the openness of the oral cavity in the vertical direction.

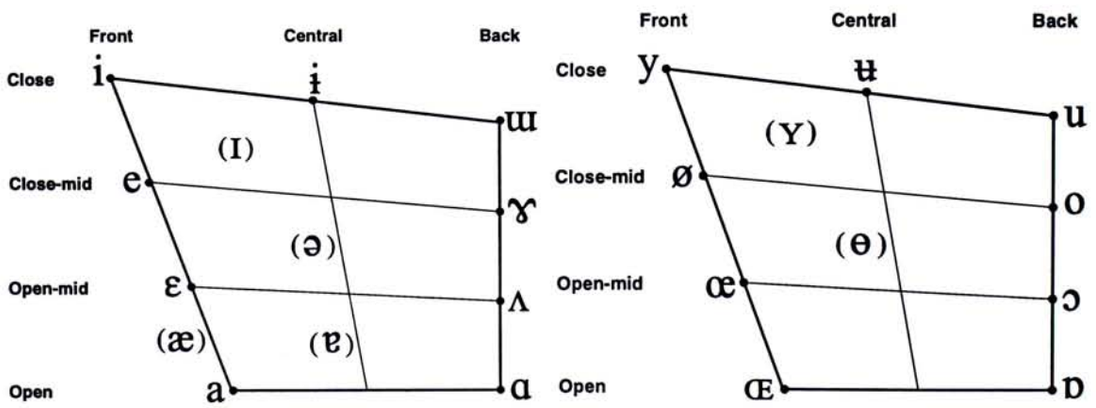

Figure 2.5: Primary and secondary cardinal vowel quadrilaterals (bracketed symbols are not cardinal vowels)

For the oral cavity openness, four theoretical positions are specified. **Close** refers to the highest position to which the tongue can achieve without producing audible friction. **Open** refers to the lowest position that a human's tongue can achieve. **Mid-open** and

Oral Cavity Openness	Tongue Position	Front		Mid		Back	
	Lip Rounded	No	Yes	No	Yes	No	Yes
Close		i (1)	y (9)	ɨ (17)	ʉ (18)	ɯ (8)	u (16)
Close-Mid		e (2)	ɸ (10)	ə	ɵ	ɤ (7)	o (15)
Open-Mid		ɛ (3)	œ (11)			ʌ (6)	ɔ (14)
Open		a (4)	Œ (12)			ɑ (5)	ɒ (13)

Table 2.6: Table of the IPA Symbols of Some Vowels. Those with numbering are cardinal vowels and the number are their numbering code.

mid-close are the two intermediate positions. This coordinate system applies to both series of cardinal vowels. In general, these cardinal vowels are numbered in sequence for reference purpose [Table 2.6]. Number 1 – 8 are used for the primary cardinal vowels and 9 – 16 are for the secondary. Later on, some phoneticians think that the highest points the centre of tongue can reach is desired, number 17 and 18 are added for the unrounded lip and lip rounded vowels. Figure 2.5 shows the primary and secondary vowel quadrilaterals together with some nearby vowels. The respective detailed articulator positions are also tabulated in table 2.6.

It is worth noting that this assignment of symbols to sounds is based on English. As specified by the International Phonetic Association, IPA [32], the closest IPA symbols is assigned to closest sounds and may not be exactly related to the same sound for the same symbols in different languages. Therefore, many deviations should be expected when different languages are considered.

On the other hand, the two separate spaces may be combined into a three dimensional vowel quadrilateral with lip-roundness as the third domain. This allows easy deployment of phonetic knowledge in the acoustic signal synthesis process. The combined quadrilateral is shown in figure 2.6. In this representation, the numbering code is not used. However, based on the mid-sagittal diagrams, coarse theoretical coordinates are assigned as shown in figure 2.6.

One of the advantages of the vowel quadrilaterals is that they assist the study of vowels. Simple vowels are treated as a particular point in this articulatory space. Any

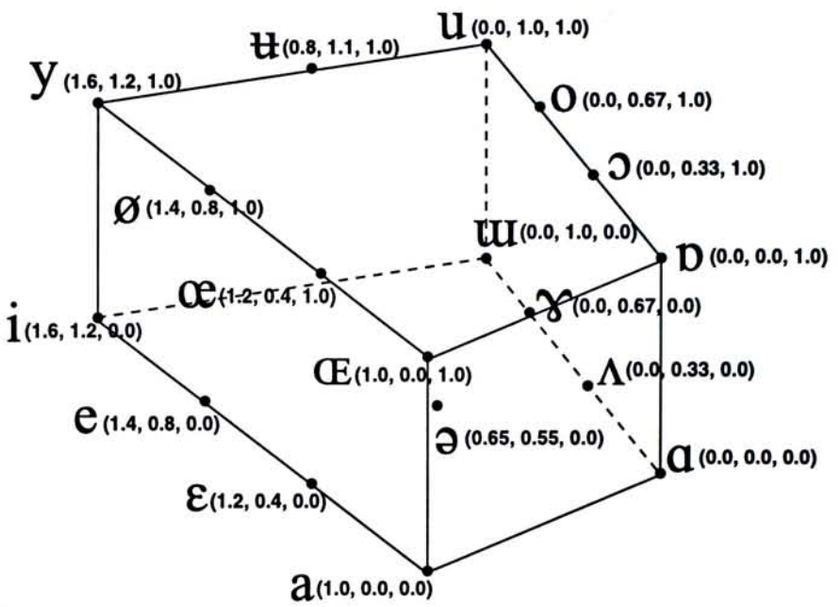

Figure 2.6: The combined cardinal vowel quadrilateral integrates the primary and secondary quadrilaterals to form the bounded three dimensional space of articulator positions.

slight deviation can be treated as coarticulatory or allophonic variations. For the study of difference between dialects or languages, this enables visual demonstration of the difference of phones as neighbouring points in the diagram. In addition, for the study of diphthongs and triphthongs, it allows these compound vowels to be treated as transition from target points to target points in the articulatory space through different paths.

Chapter 3

Speech Synthesis Technology

Speech synthesis has long been a hotly researched topics in the field of speech processing. People have tried various methods aiming at imitating the human speech production process. There have been many different methods investigated and demonstrated. Some of the pioneers built mechanical models simulating the human speech production system. While some other chose to use the electronic means. Electrical analogue of the vocal organs had been worked out by Stewart in 1922 [42]. Later, at the World Fair of 1939, the famous speech production systems "Voder" was demonstrated [43]. In addition, there were also other different synthesis systems such as the Pattern Playback by Cooper et. al. at Haskins Laboratory [44]. At that time, they tried to resynthesis speech signal from time-frequency intensity plots (the spectrogram) of original speech and also hand-made plots. Since then, people from all over the world have carried out much in depth research from different point of views for facilitating the ultimate goal of simulating human speech production.

3.1 The Human Speech Production

The human speech production process is facilitated by various organs as shown in figure 3.1. The respiratory system plays an additional role as speech production system. Air pressure from the lungs is controlled by the diaphragm and the rib-interstitial muscle. It flows through the trachea and passes through the **vocal tract**. Vocal tract is a collective term for the tube ranging from the vocal cord up to the lip. When the air-flow passes through the vocal cords, the cords may be widely opened letting air

flow through. On the other hand, they may be relaxed such that when air passes by, they will shutter and generate periodical thrust of air pressure to the vocal tract. To control the tension and status of the vocal cords, the muscles around are contracted and relaxed accordingly. Different types of vocal cord conditions will result in different types of sound sources. These sources give the speaker unvoiced and voiced sounds in the produced speech signals.

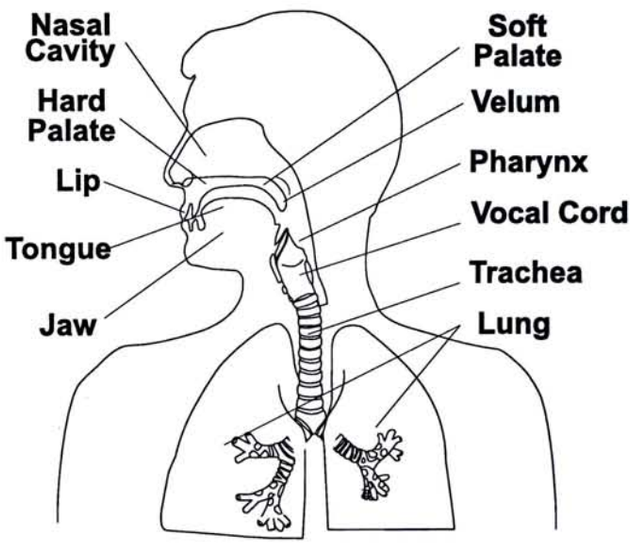

Figure 3.1: This sketch shows the main organs involved in the production of acoustic signals by human body. It consists of the lung, pharynx and various articulators. Some of these organs play the dual role for speech production and respiration.

After passing through the vocal cords, the air stream will then be modulated by the vocal tract. There are articulators that are controlled to change position so as to alter the vocal tract shape. For in this way, the acoustic properties (such as resonance, absorption) of the vocal tract are varied. Consequently, the air stream is "shaped" to the desired sound pressure pattern. Through proper control of the articulators positions, human can produce different speech signals to convey information acoustically. Some of these articulators are shown in figure 3.1.

Based on this idea, Fant [45] has proposed the source-filter model of speech production. Where phonated air stream from the vocal cords is the excitation source and the vocal tract is the filter that modulates the source signal for final output. This model is so far the most widely adopted model in speech synthesis research.

3.2 Important Issues in Speech Synthesis System

For speech synthesis technology, there are a number of issues that are either of primary concern or gaining attention [46]. Among these, the most important are : ***controllability***, ***naturalness***, ***complexity*** and the ***information storage.***

3.2.1 Controllability

Synthesis control is an important issue in speech synthesis. It concerns about the amount and degree of the synthesis control provided for external controlling agents. In order to vary the output characteristics for targeted speech signal, speech synthesis system should provide controllable system features with adequate degree of variations. In addition, the quality of the resultant output should also be taken care of.

These controls may be desired for many reasons. In real speech signal, there are as many possible variations as there are differences between different people's speech production systems and their speaking habits. In general, we need to have control in two main categories of properties: **acoustic** and **prosodic**. Acoustics includes those variations as large as for producing different words. It also includes variations as small as those in the same words but spoken at different time and under different context and conditions. It also includes the variations that allow changing the voice quality from one man to another, or even to voice of other gender, age groups or native tongue etc.

In additional to acoustic properties, prosodic properties also play important roles in the determination of the naturalness of output signal as well as the individual characteristics of speakers. Changing prosodic properties of speech signal may also achieve the goal of changing speaking rate and speaking attitude.

3.2.2 Naturalness

Naturalness is another concerned properties. Its importance is not simply for comfort, there is also psychological effect that might hinder the system's acceptability [47]. For short time interfacing, intelligible speech output is sufficient. However, if speech is supposed to be the primary interface, its naturalness is important. The user may simply become reluctant or unable to use the system because of the stress and fatigue caused

after prolonged listening to synthesized speech output. It is desirable that the speech is psychologically tolerable for prolonged usage. Currently, the exact properties in the synthesis process that determine the naturalness of output are still unclear. By the way, it is generally believed that both **acoustic** and **prosodic** characteristics play significant role.

3.2.3 Complexity

For all kinds of systems, complexity is a main concern. The tolerable complexity of a system depends on the specific application and requirements.

For speech synthesis system, there is a general trend that system with more control will be more complex. It appears in some sense that controllability and complexity are contradictive. However, natural sounding speech output can still be achieved using simple synthesis method. While complex systems do not guarantee the naturalness of speech output.

On the other hand, it is believed that the importance of this issue is decreasing as technology continues to advance. However, this is valid for large scale systems only. For small scale or even portable system, mild complexity requirement is always desirable.

3.2.4 Information Storage

Information storage concerns with the problem of what sort of information is needed and how much is required to be stored up in database for synthesis. In general, a more natural sounding synthesis system will take up more storage for templates and rules storage. When there is a need for less storage, those information not included in the database is either simply ignored or modeled by some methods. This means that when a smaller database is desired, the controllability of the method should be better. Moreover, unnaturalness may be introduced in the output due to the imperfectness of the control model.

As the computation power and storage cost are continuously decreasing, the concentration may be shifted towards how to handle large amount of information intelligently and efficiently for practical systems.

The four issues summarize the idea of many different works and assessment schemes for speech synthesis. Although they do not represent perfect evaluation scheme for the technology, they provide useful guideline for development, assessment and comparison. Until now, there is not any particular technique perform exceptionally well on all these areas. The choice for implementation is always based on the specific application and requirements.

3.3 Units for Synthesis

For the synthesis of speech signal, templates of speech signal or parameters must be stored. The units of templates are important to the resultant characteristics of the synthesis system.

The simplest way is to store up the whole sentences or paragraphs of speech and playback the appropriate one during synthesis. This templates playback method has the advantage that it has natural output and is simple and easy to use. However, large storage is required and the flexibility in using the templates is extremely small. In order to enhance the flexibility, smaller units such as phrases, words or syllables are used. These units of templates give better synthesis flexibility. As long as intelligibility is concerned, simple concatenation of word or syllable templates are still acceptable. For further improvement in flexibility, **phone** templates are adopted.

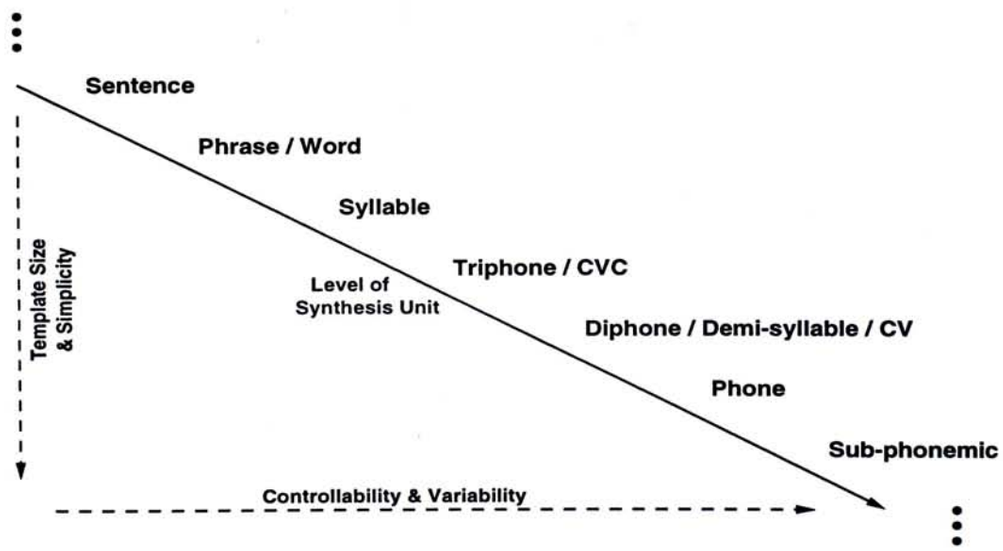

Figure 3.2: There are several levels of units common for speech synthesis. They have some resemblance to the linguistic hierarchy Fig. 2.2. In general, the lower the level of synthesis unit, the template size and simplicity decreases while the controllability increases. In order to suit different requirements, units from various levels are sometimes used together as a hybrid technique.

In general, human recognizes speech not only from the stationary acoustic properties of the vowels and consonants, but also acoustic properties in the segment transition region. Wang et. al. [48] has proposed using **diad** as the template units. They are

cut-outs of speech signal that include the transitions. They are made to start and end at nuclei centers which have stationary properties and thus better for concatenation. In fact, diad is a set of **diphones** that include various different prosodic properties. Later, Hank Truby separated the prosodic features out and called them **diphone** [49]. Diphone type synthesis has the advantage of better transition properties and the storage requirement is medium only. It was first demonstrated by Dixon and Maxey [50].

In addition, other different units have also been proposed for other properties. **Demi-syllable** [51] is similar to diphone. The difference is that diphone-cuts are made at the center of vowel nucleus while demisyllable-cut are made just after the consonant-vowel transition [52]. It reduces the problem of diphone concatenation in case diphones from slightly different vowel nuclei are joined [53].

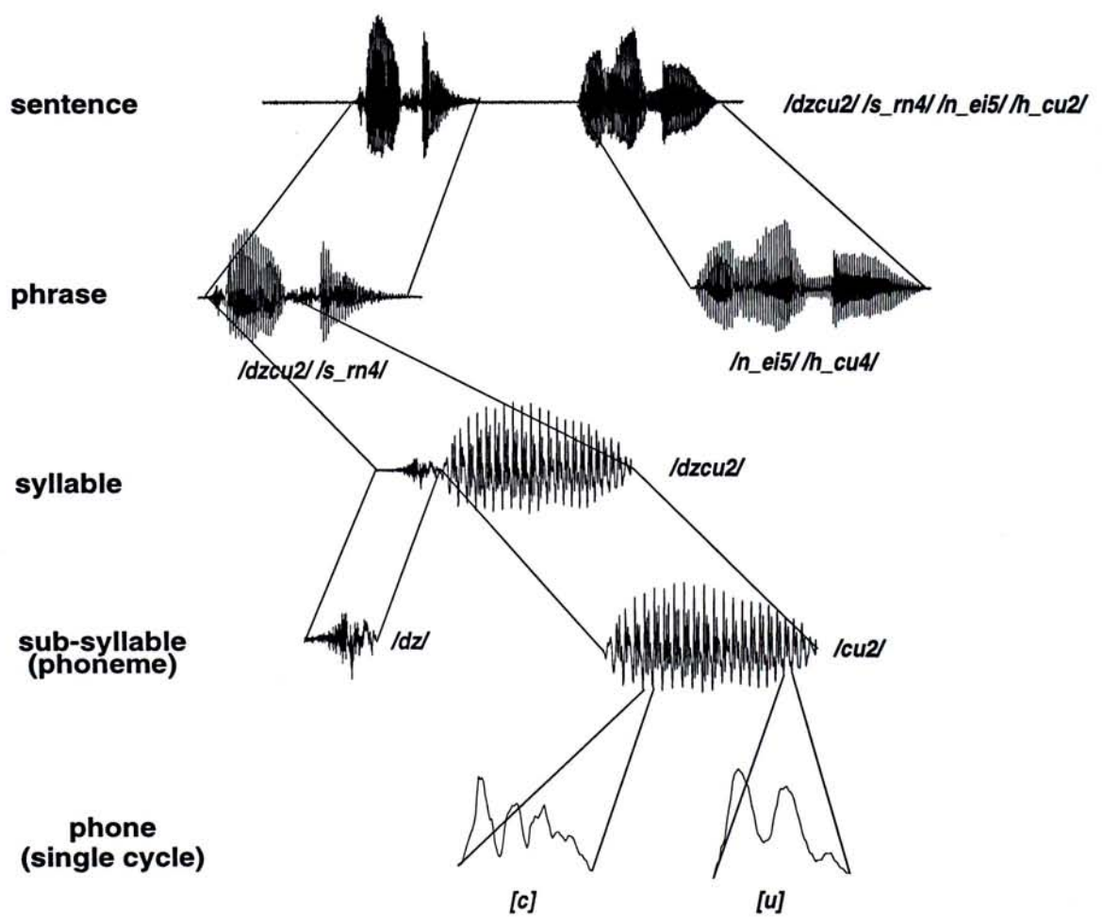

Figure 3.3: This figure shows an example of how synthesis units are made from real utterance. The speech data is cut into smaller size units as the level goes further.

Besides those linguistically based synthesis units, there are also other types of units such as sub-phonemic [54] units, triphones, hybrid type templates [55] as well as automatically selected units [3]. The selected units may be a mixture of different types depending on the specified requirement.

Figure 3.2 lists the common synthesis units. It can be compared to figure 2.2 that shows the phonological hierarchical structure of sentences. In figure 3.3, an example of synthesis units cutting is shown for some units from sentence down to single pitch cycle phone template levels.

3.4 Type of Synthesizer

There are many different types of synthesizers developed for different functions. The technique that it employed for templates storage and the method of synthesis are of primary concerns. For those information not included in the templates, rules are employed for generation of the required data. As far as the type of synthesizer is concerned, we can categorize them into three main types : copy concatenation, vocoder synthesis and articulatory synthesis.

3.4.1 Copy Concatenation

Copy concatenation method possibly has the longest history. Simple playback of recorded speech can sometimes satisfy needs of simple applications. This category may include those systems with waveform coded speech templates. Copy concatenation using higher level synthesis units [Figure 3.2] offers simplicity in the system but the controllability and flexibility are sacrificed. As more flexibility and quality is desired, people also consider concatenating templates of various synthesis units [56, 57].

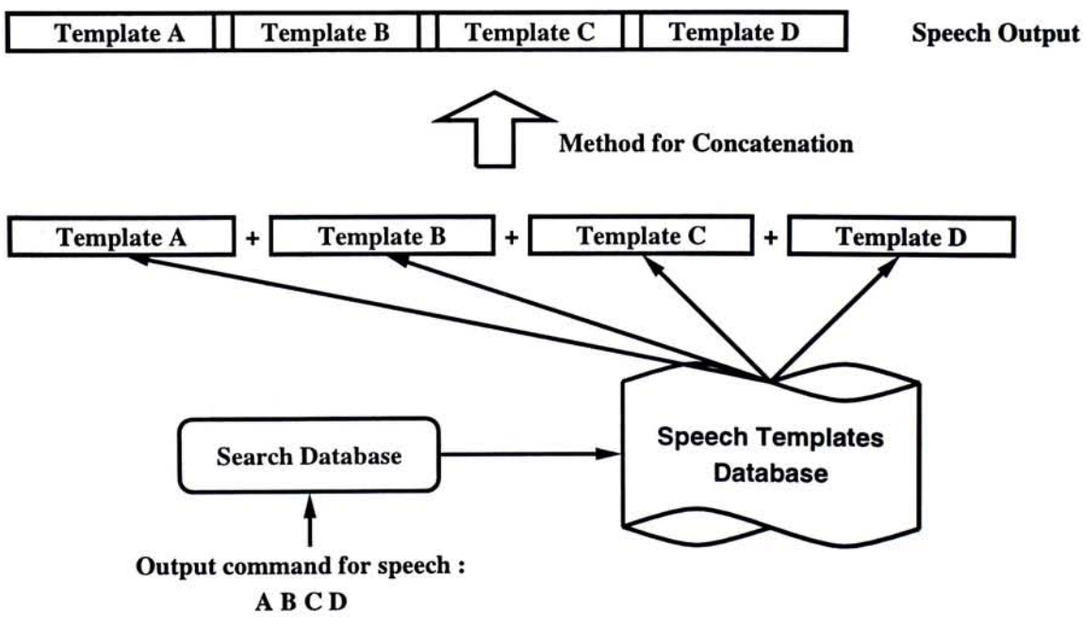

Figure 3.4: Copy concatenation synthesis is basically achieved by searching among database templates and select appropriate candidates for concatenation. The concatenation of templates may have various different strategies for more desirable joining.

This type of synthesizers usually have an inventory of recorded or coded speech templates. For synthesizing speech signals, appropriate templates are selected and concatenated [Figure 3.4]. Variations in the speech are achieved by using large template inventory or by modification techniques. Similar modification approach can also be applied to reduce discontinuity problem in template transitions.

The main advantage of this method is its simplicity and naturalness of output speech. However, as more variations are required, the storage grows larger and renders difficulty for small systems. While smaller storage can be achieved by using smaller units at the cost of additional problems (e.g. discontinuities) between template transitions.

For speech template variations, there is method proposed for speech modification that is found to be beneficial to the modification of recorded templates. The pitch synchronous overlap add (PSOLA) method [4] allows either frequency or time domain modification of the prosodic properties like speaking rate and pitch profile of the utterance under consideration. It also provides certain degree of control over the templates transitions [58]. Moreover, voice conversion and transformation has also been reported [59].

For further improvement, attention is paid to the controlled variation of stored templates and naturalness of output. Searching of templates from large database and efficient use of storage resources [60] should also be taken care. Furthermore, people also work for methods to select synthesis units automatically and efficiently from large inventory speech corpus [3].

3.4.2 Vocoder

Vocoder synthesis is another popular type of synthesis technique. It is based on the source-filter model. Vocoder type synthesizer may be made up of filter banks for modeling the vocal tract frequency response and the characteristics are controlled through the filter coefficients. Representative works include the vocoder by Dudley [61], phase vocoder of Flanagan et. al. [62] and the JSRU channel vocoder [63].

On the other hand, vocal tract characteristics may also be represented by parametric models. For example, the formant synthesizers and the family of linear predictive coding based techniques are popular ones.

Formant Synthesis

Formant synthesizer is based on modeling the resonance properties of the vocal tract. The formants represent the high energy concentration frequencies in the speech signal spectrum. Together with proper bandwidth, the spectral properties of speech signal can be represented parametrically.

Practical implementations demonstrated include the Parametric Artificial Talker (PAT) by Lawrence [64], Orator Verbis Electris (OVE) by Fant and hybrid approach by Klatt [65]. The commercial text to speech synthesizer such as DECtalk is also a formant based synthesizer.

Formant synthesizers can be implemented in different architectures, such as parallel, cascade or in hybrid approach. The parallel and cascade formant synthesizers are essentially resonant filters put in parallel and series. For any type of formant synthesizers, each formant can be digitally represented by a resonant filter section. Each resonant filter section is based on the second-order equation,

$$H_x(z) = \frac{k}{1 - 2e^{-\frac{BW_x}{F_s}} cos\frac{\pi F_x}{F_s} z^{-1} + e^{\frac{2BW_x}{F_s}} z^{-2}} \tag{3.1}$$

where $H_x(z)$ represents the transfer function of the section with formant frequency F_x and formant bandwidth BW_x. By combining several sections of these resonant filters, the speech output can be obtained by specifying formant parameters and by applying proper excitation sources to the synthesizer [Figure 3.5].

In formant synthesizers, the templates stored are the formant frequencies and bandwidths for the template speech signals. This type of synthesizer allows a great degree of flexibility by allowing control over the intensity, pitch, and timing. Template variations can be obtained by variations in formants and bandwidths for the modification or transformation of speech.

In general, rules are formulated to control the formant synthesizers parameters. Early works include the rule based formant synthesis by Kelly and Gerstman [66], the elegant table driven synthesizer of Holmes [67] and the S-transitions for formant frequencies by Mattingly [68] etc.

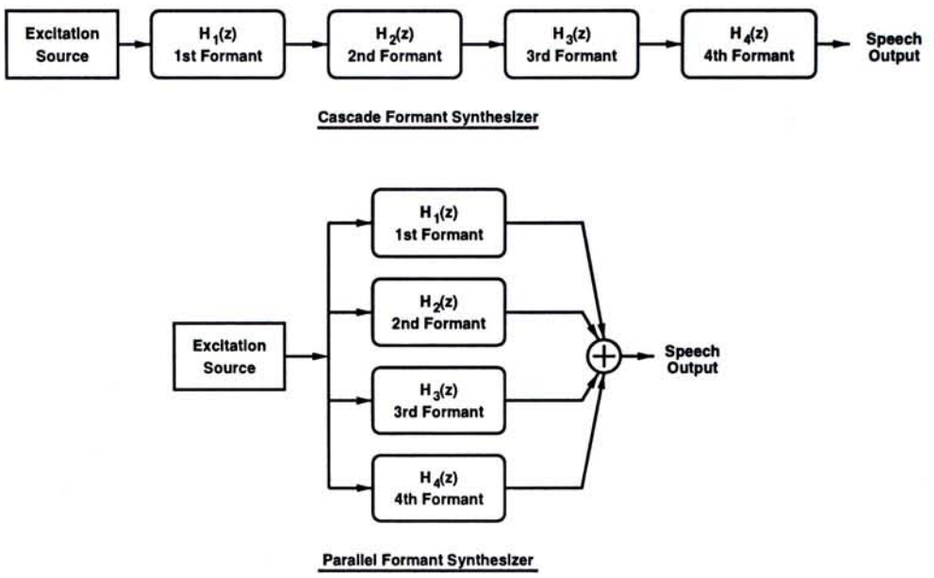

Figure 3.5: The cascade and parallel structure of formant synthesizers are shown. Each section represents the resonant filter for a formant with the corresponding bandwidth. Hybrid type can be obtained by combining both the parallel and cascade structure into one synthesizer.

Linear Predictive Coding

The linear prediction coding has imposed much impact in the field of speech processing. After the advent of this technique [69, 70, 71], many works followed in the area of coding and synthesis. The first LPC speech synthesis chip is made by Texas Instruments – the Speak and Spell [73]. It has brought the LPC based speech synthesis techniques to many practical realization [74]. In general, this technique is often used in speech coding. For speech synthesis, it is used for the parametric storage of speech templates for concatenation [Figure 3.6].

This technique is also based on the source-filter model of speech production. The LPC parameters represent the digital filter properties and the source is represented by the excitation to the LP filter. The prediction of the LPC is based on the equation

$$x(n) = \sum_{i=1}^{p} a_i x(n-i) + u(n) \tag{3.2}$$

where, x_n is the speech data, p is the prediction order, a_i are the predictor coefficient and $u(n)$ is the excitation [72].

Speech templates stored in this parametric form lack the desired controllability

through simple manipulation of the LP parameters. To enhance the properties of the stored parameters, representations such as line spectrum pair (LSP) and partial correlation parameters (PARCOR) are often used.

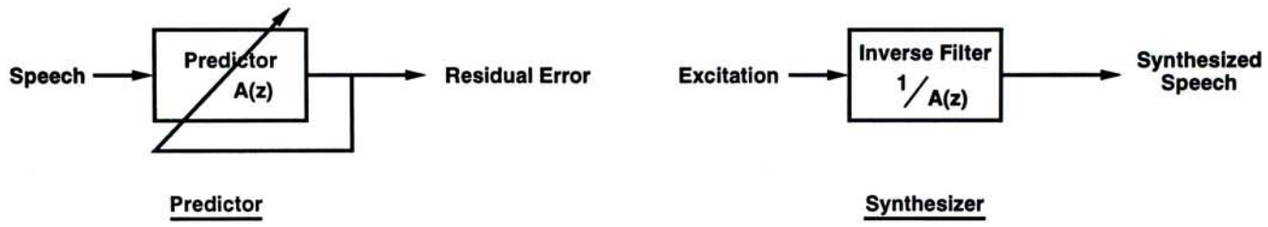

Figure 3.6: The predictor and synthesizer for the LPC based synthesis technique are shown. The predictor coefficients are usually obtained for a frame of speech data. Excitation is then applied to the the inverse filter for synthesizing the speech signal.

3.4.3 Articulatory Synthesis

Besides the direct copy concatenation and the parametric models, there are also intuitive synthesis technique that has been investigated – **Articulatory Synthesis**. The idea behind articulatory synthesis is that with detailed modeling of the vocal tract cross-sectional area profile, the vocal tract wall properties and also the glottal source etc., human speech production can be reliably simulated. However, the complexity, lack of knowledge and the high cost still hinder easy achievement of this ideal.

Articulatory synthesis tries to model the human speech production process and provides controls with correspondence to real articulators. There has long been interest in this type of synthesizer and many related works have been carried out [75, 76, 77, 78, 79]. The first English text articulatory synthesizers are made by Teranishsi et. al. [80] and Matsui [81] and representative works in articulatory synthesizer include [82, 83, 84, 85].

In articulatory synthesis system [Figure 3.7], an excitation source for the speech production is included. For speech production, it is actually the air stream passing through a gap between the vocal cords. For making reliable excitation source, there have been many models proposed for the vocal cords [86, 87].

On the other hand, there are also many different articulatory models proposed for the vocal tract. Ranging from simplified models [82] to more complicated vocal tract

area functions [88, 89]. Nasal tract [90] and lip radiation [91] are also incorporated for enhancing the model. In order to obtain better models, people also tried to include loss in the vocal tract wall [92], take account of varying vocal tract length and model the interaction between the glottal source and the vocal tract.

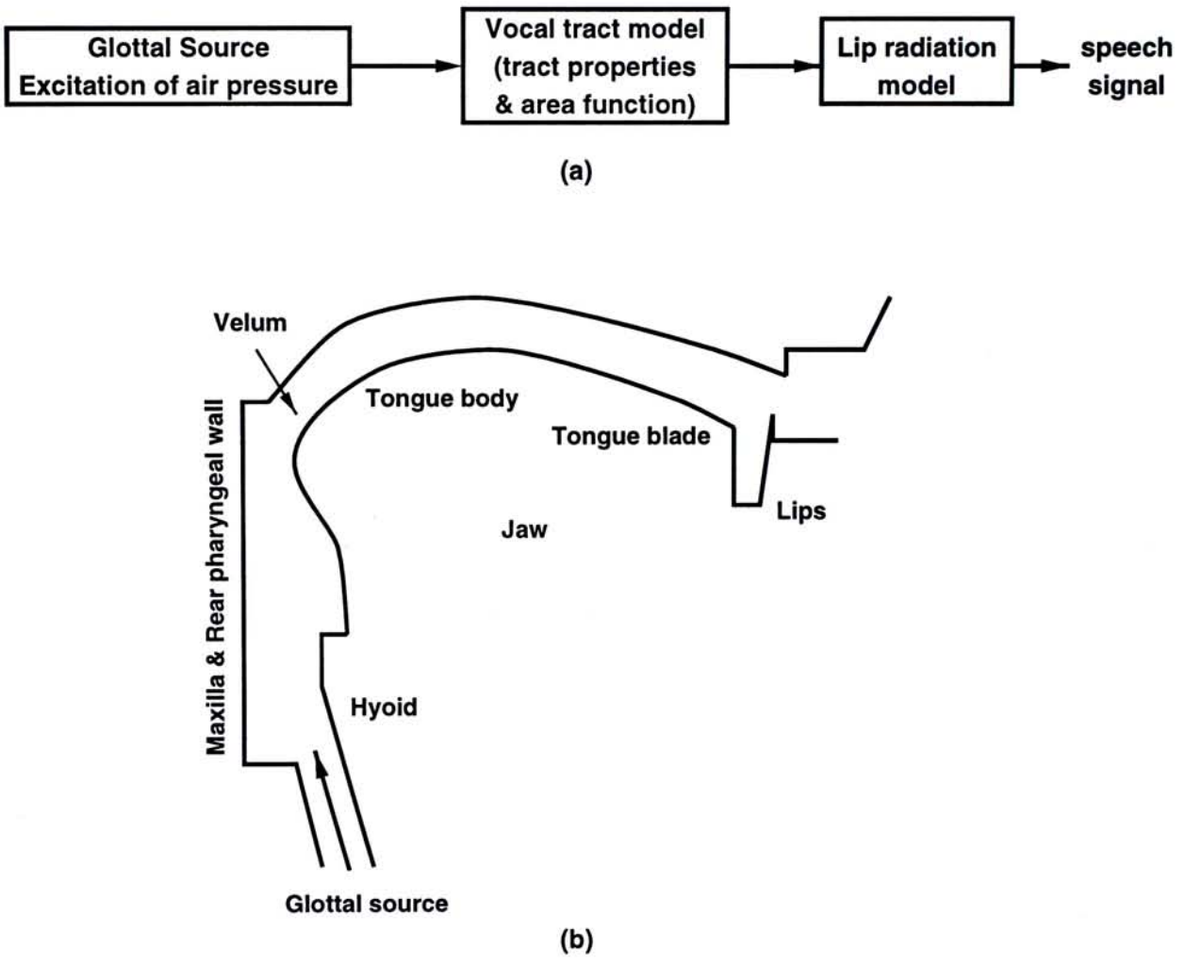

Figure 3.7: Figure (a) shows the general idea of articulatory synthesis. Glottal excitation generates source signal that is modulated by the vocal tract model and lip radiation to produce speech signal. Figure (b) is an example of the mid-sagittal vocal tract model (lossy wall and source interaction not included). (Simplified from P. Mermelstein)

This technique is generally believed to have the greatest potential as the ultimate solution for speech synthesis. The modeling of actual speech production process gives it much flexibility and room for improvement. Main problems of this type of synthesizer are the lack of synthesis rules, complexity of reliable model and the difficulty of the inverse problem.

For the inverse problem of estimating vocal tract information from speech signal, many different works have also been done [93, 94, 21, 95]. The extracted information enables easier formulation and verification of the controlling rule for the articulatory synthesizers.

In figure 3.8, the simplified relationship between various types of synthesizer is illustrated and serves as a conclusion for the description of their characteristics.

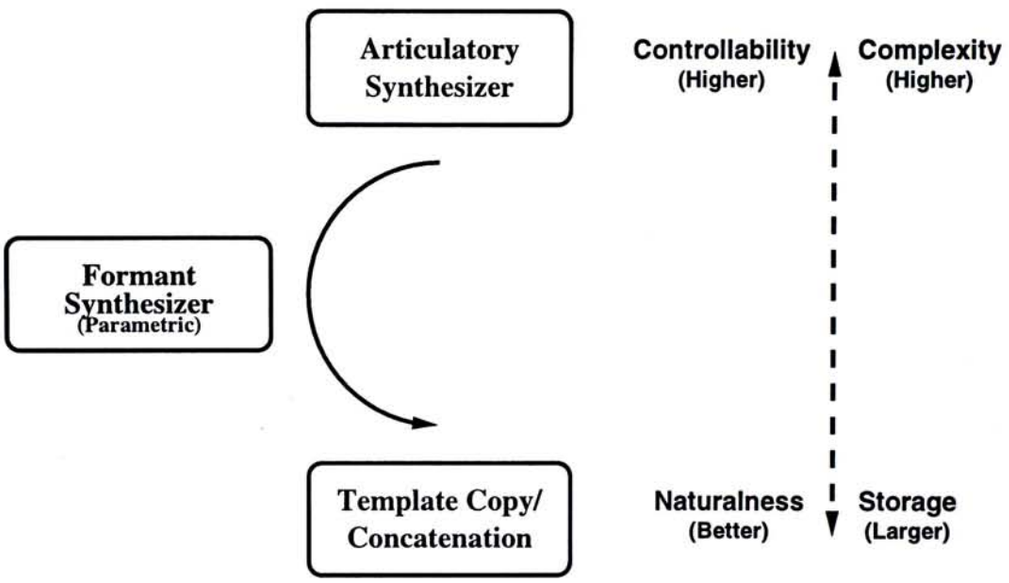

Figure 3.8: This figure shows the main characteristics of the three main type of synthesizers and the relation among them. In general, the complexity, naturalness, controllability and information storage involved are to be balanced for particular area of application.

Chapter 4

Neural Network Speech Synthesis with Articulatory Control

In this chapter, we will discuss the use of artificial neural network (ANN) for the synthesis of speech signal. Neural network speech synthesis provides an alternative solution that is complementary to the existing techniques. In this method, spectral data of recorded speech signals are used as the training templates. Synthetic utterances are obtained by providing proper articulatory input control to the neural network. It will estimate the desired output based on the training information. Exact templates will be returned if available and approximation is made otherwise. Hence, it is essentially a non-linear approximation of the training data for speech synthesis purpose.

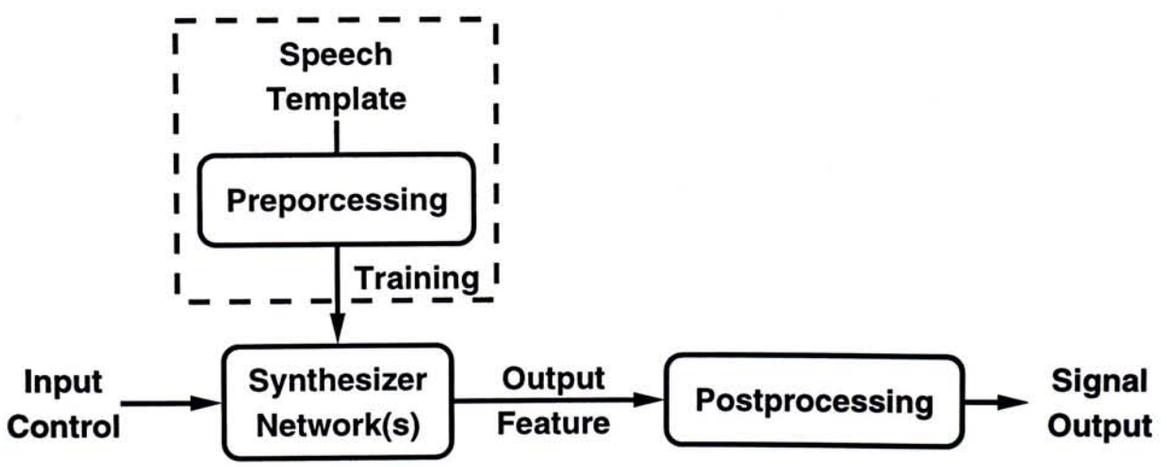

Figure 4.1: The block diagram illustrates the general components in a neural network speech synthesizer.

In figure 4.1, the block diagram illustrates the idea of speech synthesis using neural network. In general, the input control and the output feature can be selected based on application and requirements. In this work, the output feature is chosen to be the spectral information and the input control is some simplified articulatory parameters aiming at more control flexibility.

4.1 Neural Network Approximation

Neural network is a powerful tool for the approximation of complicated relationship through learning. In general, the mapping of input-output relation from examples can be formulated as an approximation problem which aims at finding an approximation for the target relationship.

For approximation, methods such as regression, shaping and fitting by polynomial, Bezier or spline have been widely applied. The use of neural network for approximation and interpolation gives flexibility in the architecture and capability for high complexity and non-linearity in the approximation. Moreover, classical approximation methods can be formulated as particular cases of the network approximation [96].

4.1.1 The Approximation Problem

The goal here is to find a function $F(W, X)$ that approximates a targeted function $Z = f(X)$, $\{f : X \in \Re^m \rightarrow Z \in \Re^n\}$. This can be achieved from a priori knowledge of the function $f(X)$ and make up the approximate function $F(W, X)$. On the other hand, it can be done through learning from a set of typical input-output pairs, $\{X_i, Z_i\}$. The learning method is useful in cases where only little knowledge of the function $f(X)$ is available or only specific domains of the function $f(X)$ are needed to be approximated.

In order to obtain approximation through learning, there should be sufficient sample data available for training. These examples should at least include representatives from the domains of interest. If the sample data set includes certain degree of redundancy, better generalization result of the input-output relationship can be achieved. Proper degree of redundancy in sample data not only improves generalization, but also makes the resultant approximation more tolerable to noise in data. In general, generalization

of input-output relationship is based on data redundancy and smoothness assumption across data samples.

4.1.2 Network Approach for Approximation

Network approach can be used to solve the approximation problem. The network is trained with sample data to approximate the targeted input-output mapping. The mapping may have continuous smooth values rather than discrete clusters as in the classification problem [96].

Neural network approximation is equivalent to extracting the desired target relationship using a composition of basis functions to provide higher degree of complexity and non-linearity in the approximation process. In actual implementation, classification and approximation networks can be viewed as the application of same method to different problems with different goals. In a classification network, "teacher" data usually have values near the extreme sides of activation functions. The "teacher" data in an approximation network, however, are not necessarily at the extreme sides of the activation function (even exists) unless it is a known property in the target mapping. In general, continuous and smooth mapping is preferred in approximation networks while clustering is preferred in classification networks for decision making.

On the other hand, both approximation and classification networks may suffer from training problems such as local minima or poor convergence. There may also be problem of over-fitting and oscillation in the generalized regions. Putting too much resource (large network size) may lead to over-fitting and sacrifice generalization for higher accuracy while smaller network will have smoother generalization. However, too much smoothing will lose accuracy in memorized data. It is necessary to strive a balance between accuracy in training and generalization for the application.

Basis Function and Activation Function

In the general network formulation, we have

$$Z = F(u(W, X)) \tag{4.1}$$

where Z is the output, $F(u)$ is the activation function and $u(W, X)$ is the basis function.

There are many different basis functions applicable for network approximation. The two main categories are linear basis function (LBF) and radial basis function (RBF).

Linear Basis Function : The output, u, is a linear combination of inputs. In multi-dimensional cases, this formulation of basis function will form hyper-planes for the network approximation. An example first order linear basis function with i neurons and n inputs is given by

$$u_i(W, X) = \sum_{j=1}^{n} w_{ij} x_j \tag{4.2}$$

Radial Basis Function : The output, u, is given by a non-linear (e.g. second order) combination of the input values. In the multi-dimensional cases, this formulation of basis function will form hyper-ellipses for the network approximation. An example of radial basis function is given as

$$u_i(W, X) = \sum_{j=1}^{n} w_{ij} \sqrt{(x_j^2 - c_j^2)} \tag{4.3}$$

with function centers c_j. For symmetric weight matrix W, the formulation will be reduced to the case of hyper-sphere in the approximation.

The activation functions, on the other hand, are usually used in amplification and non-linear warping. They determine the properties of the corresponding neurons.

Linear Function : Linear amplification is the simplest form of activation function. It is commonly used for output neurons of networks that require unbounded output range. It is defined as

$$u(x) = wx + \theta \tag{4.4}$$

where w is the amplification factor and θ is the offset.

Hard Limiter : This is a simple non-linear activation function defined as

$$u(x) = \begin{cases} -w & : \quad x + \theta < c \\ w & : \quad x + \theta > c \\ 0 & : \quad x + \theta = c \end{cases} \tag{4.5}$$

with amplification w, offset θ and c is the center of the function. A sharp, two-valued output similar to crisp logic is obtained. However, this function is not often used for its differential discontinuity at $x = c$ which makes it difficult for training and analysis.

Sigmoid : The sigmoid function is defined as

$$u(x) = \frac{w}{1 - e^{k(x-c)}} \tag{4.6}$$

where w is the gain, c is the center (threshold) and k is the crispness (temperature). Its differentiability makes the training and analysis simpler by making the deployment of backpropagation training algorithm easier.

Linear and Non-linear Approximation Network

By combining different basis and activation functions, we can get different approximation network which can be linear or non-linear in the approximation.

Linear Approximation Network A simple way of approximation is through the optimization of a least square error function. Widrow [97] has formulated the network model for the linear least square approximation. The Adaptive Linear Element (ADALINE) [97] is described by

$$Y = W \cdot X \tag{4.7}$$

where optimization is done by minimizing the cost function

$$E = \frac{1}{2} \sum_{m=1}^{M} \left(y^{(m)} - t^{(m)} \right)^2 \tag{4.8}$$

with X and W being N dimensional vectors of input and network parameters. M is the number of training data and $t^{(m)}$ the targeted training output. Formulation in this approach will result in a single layer network for the problem. This is because no matter how many layer the initial formulation is, it can eventually be simplified as a single layer linear network mathematically. Whilst, this linear approach also has the advantage that a **globally optimal** solution is generally available unless in the event that the training data makes the linear system become singular.

This type of network is usually used to continuously updating data which is useful for the estimation of input-output relation in real-time linear systems. General concern of ADALINE is the speed of convergence in the adaptive process. For approximating complex non-linear systems, this approach may not be suitable.

Non-linear Approximation Network Non-linear approximation network is useful for approximating more complex systems although model parameters are sometimes more difficult to estimate. To incorporate the non-linearity, it may be achieved by piecewise polynomial or spline at defined order of differential continuity. Non-linearity may also be obtained by nested functions and employment of non-linear basis and activation functions.

In a two layer network, using suitable basis function will arrive at classical approximation schemes in a network approach [96].

$$F(W, X) = \sum_{i=1}^{m} W_i \Phi_i(X) \tag{4.9}$$

Through the selection of $\Phi_i(X)$, we can formulate the classical approximation functions like splines and many orthogonal basis as approximation network. As an example, if $\Phi_i(X)$ is chosen to be powers and products of input X, the result is a polynomial approximation.

In general, the nested multilayer network can be described by

$$F(W, X) = \sigma_1 \left(\sum_{n_1} w_{1,n_1} \sigma_2 \left(\cdots \sum_{n_{M-1}} w_{M-1,n_{M-1}} \sigma_M \left(w_{M,n_M} X \right) \cdots \right) \right) \tag{4.10}$$

with $\sigma(\cdot)$ being the activation function. For a two layer network, it becomes

$$F(W, X) = \sigma \left(\sum_{n_1} w_{1,n_1} \sigma \left(w_{2,n_2} X \right) \right). \tag{4.11}$$

With a linear basis function at input layer, the activation function then becomes the basis of approximation for the following layer. This relation is repeated until the final output layer is reached. If linear activation function is used, the formulation will degenerate to a linear network. Although this recursive basis function approach is rare in classical approximation problem, it has high potential capability in many difficult function approximation problems in network approach.

4.2 Artificial Neural Network for Phone-based Speech Synthesis

In speech synthesis, the ideal solution is faithfully imitate the speech production mechanism. However, this approach is still far from perfect because of its complexity in deriving a reliable model and an incomplete knowledge about the complete mechanism. From an estimation point of view, linear parametric models can be adopted which treat this as a system identification problem. The formant synthesizer and the LPC family of digital filter models are examples that have gained much attention. On the other hand, for producing natural speech output, record playback or copy concatenation are simple solutions. However, they have drawbacks that good quality and naturalness are usually associated with large synthesis units. These large units also lack flexibility and controllability. If small units are used to save storage, improve flexibility and simplify the template-based searching, problems will occur in the modeling of template transitions and property variations.

In order to provide a method for phone-based concatenation synthesis that allows flexible input control, neural network is regarded as a viable alternate solution [98]. Small templates database is inherently associated with the phone-based approach. While, flexibility for synthesis control is achieved through the manipulation of a set of articulatory input parameters that are fed to the neural network. This method provides a basis for comparable control flexibility as articulatory synthesis. During synthesis, transitions between templates are approximated by the neural network. It will reduce undesirable effect due to template discontinuities and insufficiency of acoustic variations in training templates. This method attempts to integrate various nice properties from different synthesis techniques into a single approach for the production of speech signals.

4.2.1 Network Approximation for Speech Signal Synthesis

Network approximation is chosen in this method for the reason that it is, at least, theoretically capable of mapping arbitrary complex relation between input and output spaces. It has been proven that the network approach can approximate arbitrary complex mapping provided that the network size and training data are adequate [99, 100, 101]. In

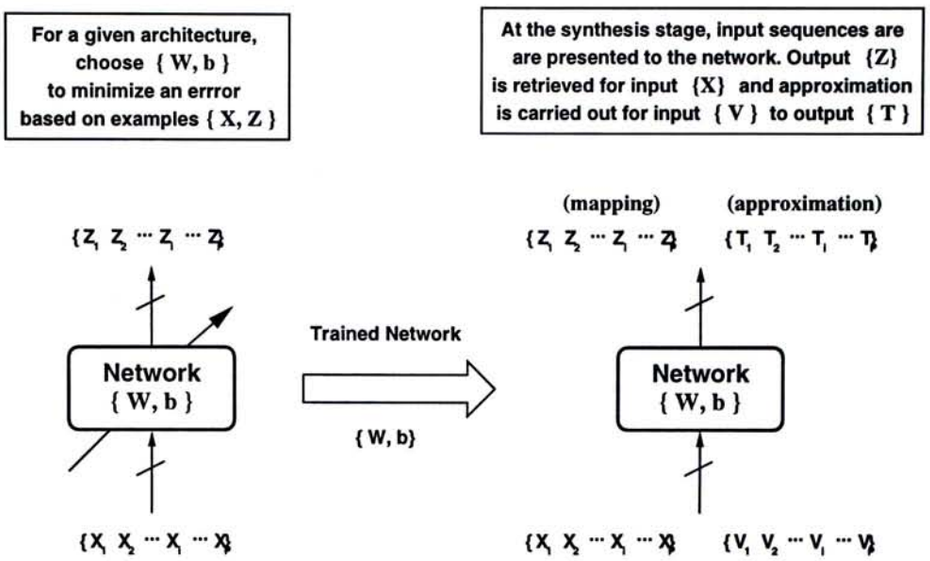

Figure 4.2: During the training stage, speech phone templates are used to train the neural networks as synthesizer network for speech synthesis. For synthesizing output, control parameters are fed to the inputs of the synthesizer network. The desired speech signal will be obtained after the post-processing stage.

speech synthesis, neural network first stores up the training templates for later retrieval. The accuracy of stored templates is determined by the provided network resource. The neural network can also provide non-linear approximation for input-output mapping in the untrained spaces. Hence, it is used to model the vocal tract of human speech production system through proper training.

The spectral approximation problem for speech signal is formulated as a mapping between the input control parameters and the spectral information,

$$Z = f(X, W) \quad \begin{cases} X \in \Re^m \\ Z \in \Re^n \end{cases} \tag{4.12}$$

where X is the input control parameter vector, Z is the output vector of spectral information and m, n are the dimension of the input and output spaces respectively. In this case, it is to find a function, $\{f : R^m \rightarrow R^n\}$, with network parameters W, that can approximate the real mapping as good as possible. By training the network using data set $\{X, Z\}_i$, we can estimate the network parameters, W, that optimize some preset objectives. Upon obtaining the network parameters, the synthesis process is facilitated by controlling the input vectors.

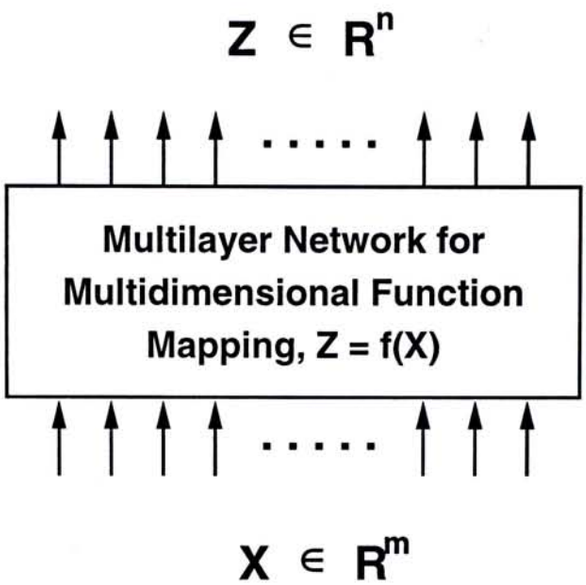

Figure 4.3: The multi-dimension function approximation problem is formulated as a network problem.

If the network is trained with sample spectral templates, it will allow a reliable retrieval of spectral information at the training locations. Therefore, more templates will result in more reliable modeling of the vocal tract properties. Depending on the architecture chosen, the spread of the training templates may affect the ultimate result. The space converge of the templates should be large enough to cover most of the occupied articulatory space. After all, this approach of network approximation through learning can incorporate information and knowledge of the vocal tract spectral properties that is either known or unknown into a single network solution.

In the synthesis process, for any input vector which may be trained or untrained, the synthesizer network will return the approximated spectral information as its output. This approximation is achieved through certain degree of redundancy in the set of training templates. Moreover, smoothness is assumed among the training templates. This approximation of spectral information is useful for characterizing the transition between templates during synthesis to avoid any discontinuity. It is also useful for the modeling of any variations of the output due to either allophonic or coarticulatory variations. These variations are simulated by corresponding changes in the input vectors controlling the synthesizer network.

Another property of the network approach for spectral information approximation is that it gives a non-parametric model for the vocal tract spectral information. In addition, training the network with vowel templates will simplify the source-filter interaction and so can reliably reproduce the spectral content. Modeling of different untrained types of sources are achieved by applying manipulating the spectral envelopes (In the form of the phase properties and spectral envelope changes).

4.2.2 Feed forward Backpropagation Neural Network

The simplest choice of network architecture for approximation is the linear network. However, optimization of the linear network is simply a least mean square fitting for the training data. The resultant network is linear and the approximation is just a linear interpolation. This is obviously insufficient for the vocal tract model.

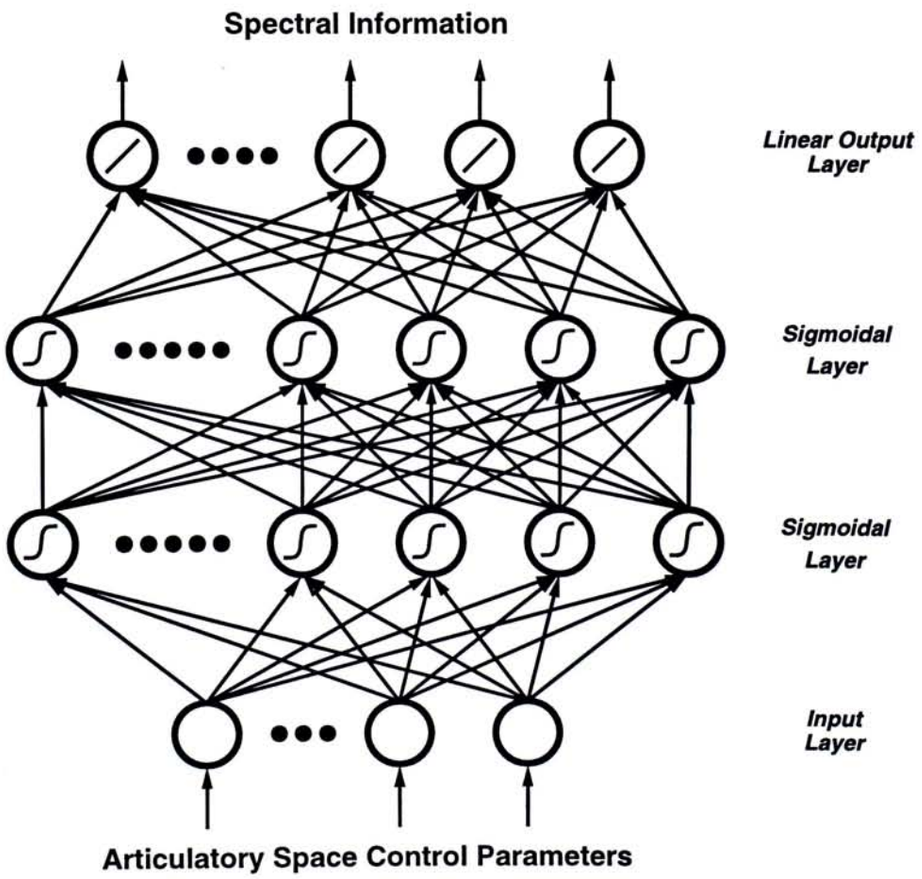

Figure 4.4: The multi-layer feed forward backpropagation network used for the speech synthesis problem

In order to achieve a more appropriate model, non-linear elements are incorporated in the approximation process. Multi-layer perceptrons with non-linear activation functions are possible choice. In these feed forward networks, non-linearity can be incorporated through the use of activation functions. Sigmoid functions are usually applied since it enables adoption of the error backpropagation training.

A single layer feed forward backpropagation network is actually a linear basis approximation function with a sigmoidal activation function. When more layers are added, the approximation becomes more complicated. The linear basis approximation with sigmoidal activation function is cascaded again and again til the output layer. The output from the activation function of previous layer is treated as the input to another linear basis approximation with sigmoidal activation function. For a non-trivial network size, this will create highly complex and non-linear basis for the approximation problem. Additional linear output layer can also be applied to provide extra mapping of the network approximation result to an unbounded space.

By using the feed forward backpropagation network for speech synthesis, reliable spectral outputs for corresponding input vectors can be achieved for those within the training templates. On the other hand, for input vectors not lying within the training set, approximation is carried out. Based on the information in the training templates, a smoothing process is performed to obtain the resultant spectral output. Therefore, the accuracy and generalization of the network rely on the amount and coverage of the training data. More templates and wider coverage of the input space will give better approximation over the whole space at the cost of a larger and more complicated network.

If the training templates are concentrated around certain space rather than evenly spread, the approximation may tend to be biased towards those specific domains. In order to overcome these problems, there are several different possible solutions such as to collect data statistically and to perform input vector quantization. On the other hand, we may pre-bias the training data based on statistical results so as to equalize its effect. Lastly, we may also simply avoid undesirable clustering of the training templates through a careful templates selection process that bases on linguistic knowledge of the phone distribution.

4.2.3 Radial Basis Function Network

Radial basis function has long been applied to tackle the approximation problem in many different engineering applications. The radial basis function approximation can also be formulated in normal network approach.

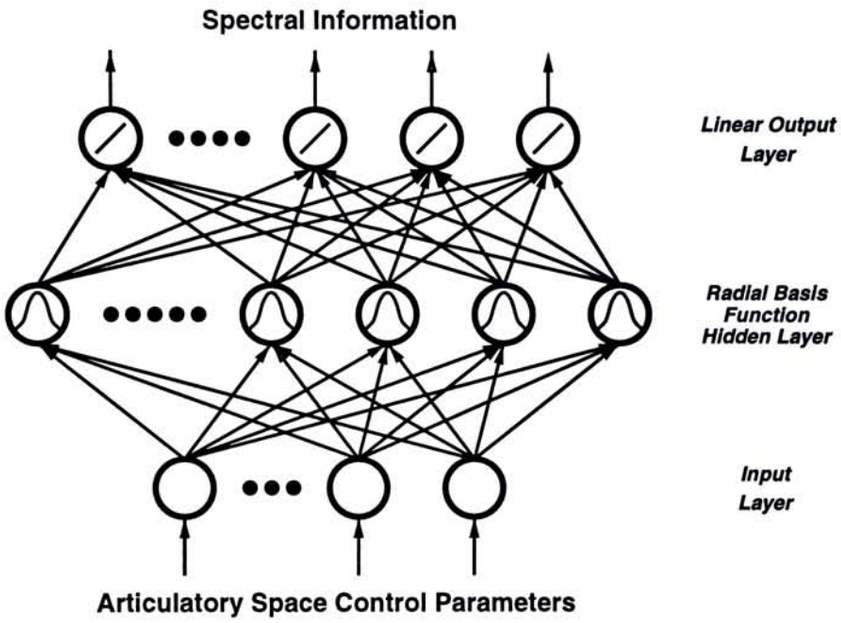

Figure 4.5: The Gaussian radial basis function network for the speech synthesis problem.

There are many different radial basis functions available for use as interpolation basis. The Gaussian radial basis function is a typical example. It has been shown that [96, 102] approximation using the Gaussian radial basis function is a particular solution of the least square error formulation with regularization on the derivatives of the basis function. The energy function to be minimized is

$$H = \sum_{i=1}^{N} (t_i - \phi(x_i, w))^2 + \|P\phi\|^2 \tag{4.13}$$

where $|\cdot|$ denotes the L^2 norm and $P(\cdot)$ is the constrain operator.

If the constraint operator, P, is chosen to be the second order derivative of the function ϕ, the resultant basis ϕ function is a Gaussian function [96, 102].

$$\phi(x) = \sum_{i=1}^{N} w_i e^{-\frac{\|x-c_i\|^2}{2\sigma^2}} \tag{4.14}$$

Applying this approximation to speech synthesis, we will arrive at a multidimensional basis function, ϕ. Under predefined spread σ, the centers c_i are chosen and the weight w_i are optimized. The result is a network with an input layer, a hidden Gaussian radial basis function layer and an output layer which is a linear combiner for the basis functions.

4.2.4 Parallel Operating Synthesizer Networks

In this neural network approach for speech synthesis, architectural simplification can be achieved by using parallel operating synthesizer networks. It can reduce the complexity of each synthesizer network by reducing the dimension of the controlling space for each network. The use of parallel networks should be made properly by taking into account the actual articulation process for producing the sound. Different parallel synthesizer networks are used to produce different types of sounds. The adoption of parallel synthesizer networks provides simplification in the synthesizer architecture and avoid the training of large neural networks.

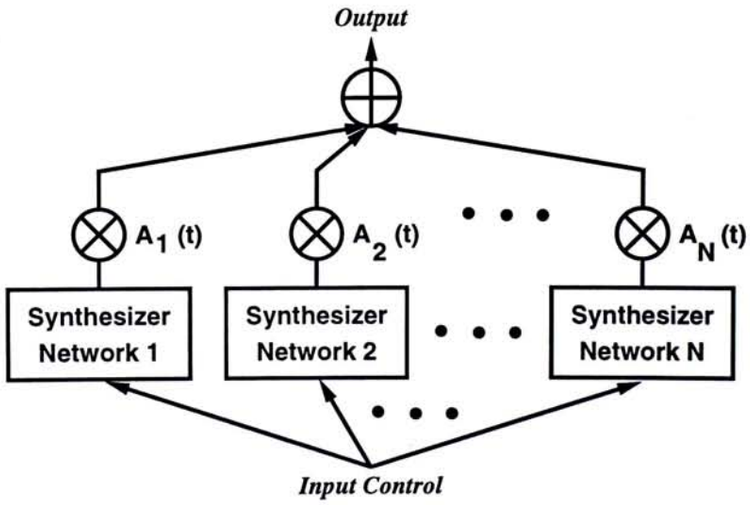

Figure 4.6: Parallel synthesizer networks are used for simplifying the overall synthesizer architecture.

Since the adoption of parallel networks is a simplification in the overall architecture, it will bring about certain degree of degradation in quality. In order to minimize this effect, the use of parallel synthesizer networks are usually limited to cases where the transitions across phonetic segments are so short that perceptual change of the output

quality will not be noticeable. From the articulatory phonetics point of view, this simplification should be applied to those segmental transitions that involve fast flip-flopping of articulators. In Cantonese, transition to and from nasal consonants is an example.

In figure 4.6, a block diagram is shown to illustrate the general idea of the architectural simplification. If parallel networks are not used, all of the control parameters must be integrated into a single synthesizer network and the dimension of the control space, the size if the synthesizer network size will become very large. This may not be desirable if the increase in complexity is far more than the gain in controllability and quality.

4.3 Template Storage and Control for the Synthesizer Network

The proposed technique of using neural network for speech synthesis has two important characteristics : firstly implicit storage of the templates information as network parameters, and secondly, a selectable input control for the synthesizer network. This method does not require additional external template database. Moreover, information is implicitly stored as network parameters. This not only allows more compact storage, but also provides a basis for the approximation of the template transitions. On the other hand, it allows user designed input control for the synthesis process. Articulatory parameters are selected so as to enable more intuitive and flexible control. Modeling for various kinds of variations in speech signals are thus simplified.

4.3.1 Implicit Template Storage

In this approach, the information in the training templates are stored implicitly as network parameters through training process. This allows the storage capacity to be set by configuring the network size. The data to be stored as the network parameters will be automatically selected by the training process. In general, larger network will store the information in a more reliable sense. Although smaller network is also feasible, it is done at the cost of losing some fine details and hence the accuracy is lowered.

Network training will also take advantage of the redundancy in the training information for estimating the network parameters. It will make use of the redundancy for information generalization during the approximation process. The success of network approximation is believed to rely on certain degree of redundancy assumed in the training data provided [96]. The choice of network model size will determine the generalization properties. While Over-modeling will cause oscillation, under-modeling will lose accuracy during the reproduction of the templates.

4.3.2 Articulatory Control Parameters

The input control to the synthesizer network is articulatory based so that a more intuitive and flexible control for synthesis can be achieved. There have already been a fair

amount of works in articulatory synthesis [88]. In this proposed synthesizer network, the articulatory parameters are chosen based on the combined cardinal vowel diagrams. Three parameters are included and they are

- **Oral Cavity Openness,**
- **Tongue Body Front-end Position, and**
- **Lip Roundness**

which constitute a three dimensional space of input control.

Table 4.1: The table lists out the relationship between the formant frequencies and the characteristics of the vocal tract for speech signal [103]

Length Rule : The average frequencies of the vowel formants are inversely proportional to the pharyngeal-oral tract length. The longer the tract length is, the lower the average formant frequencies are.
F_1 **Rule – Oral Constriction :** The frequency of F_1 is lowered by any constriction in the front half of the oral cavity. Greater constriction implies a lower F_1 in the speech sound.
F_1 **Rule – Pharyngeal Constriction :** Constriction in the pharynx will raise the F_1 frequency. The greater the constriction, the more the F_1 is raised.
F_2 **Rule – Back Tongue Constriction :** The formant frequency F_2 tends to be lowered by a back tongue constriction. The greater the constriction, the more F_2 is lowered.
F_2 **Rule – Front Tongue Constriction :** The frequency of F_2 is raised by a front tongue constriction. The greater the constriction, the more F_2 is raised.
Lip-Rounding Rule : The frequencies of all formants are lowered by lip-rounding. The more the rounding, the more the constriction and subsequently the more the formants are lowered.

The first two dimensions make up the basis of the cardinal vowel diagrams. They are sufficient for the specification of some vowels in common languages. The additional lip-roundness constitutes the additional dimension for differentiating lip rounded and

unrounded vowels. By combining the three dimensions together for vowel specification, we have defined a three dimensional hyper-space for speech sounds. In fact, higher dimensional control space may be used for the the synthesizer network. However, the higher the dimension of the control space, the more the amount of training templates required for more accurate model for the vocal tract. In practice, the choice of the control dimension and the parameters is subjected to application and performance requirement.

The use of articulatory space parameters as input control has a very good reason. Although the formants of speech signal are not simple linearly related, they are found to be related to the articulator position during the production process [Table 4.1]. Figure 4.7 shows the clustering of the vowels in the F_1–F_2 space. Comparing with the vowel coverage of Cantonese speech in figure 4.8, we can see a crude relation between spectral variation and the articulators position.

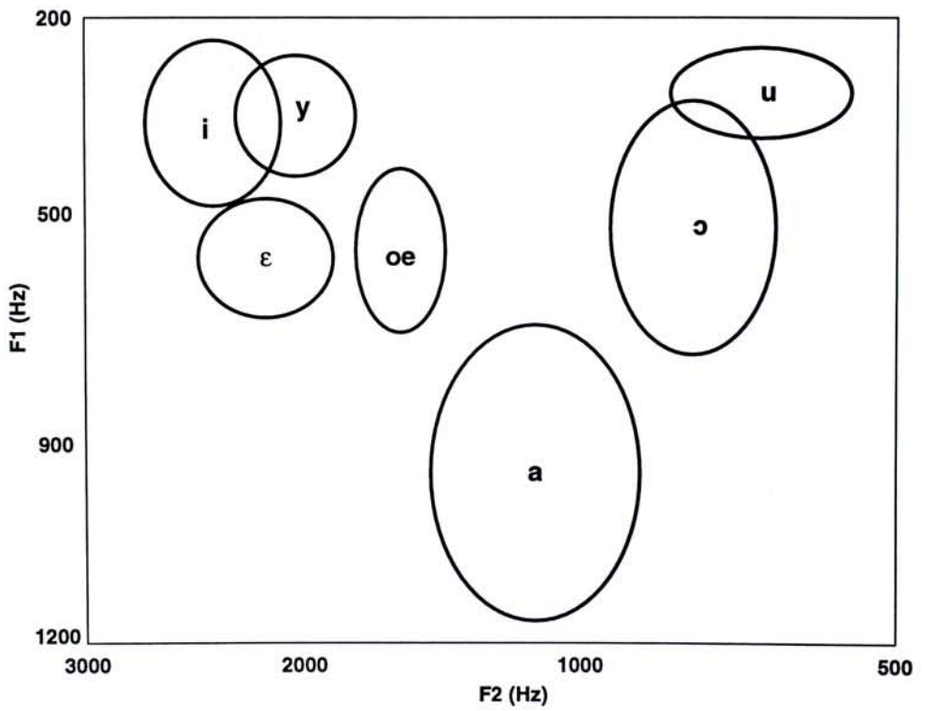

Figure 4.7: The clustering of the vowels in Cantonese consonant-vowel syllable is shown. Obvious resemblance of the articulatory space for vowels can be observed. [40]

However, the trends shown do not imply any specific detail of the spectral transition trajectories. Some traditional methods rely on detailed rules for these spectral transition. With the articulatory parameters controlled synthesizer network, these rules are

incorporated into the vocal tract network model through the training process. During synthesis, only the path of the articulatory control parameters are needed for the reproduction of speech signal. The complex spectral transitions are retrieved through the trained network mapping and problems concerning the rules of parameters trajectories are transformed and simplified.

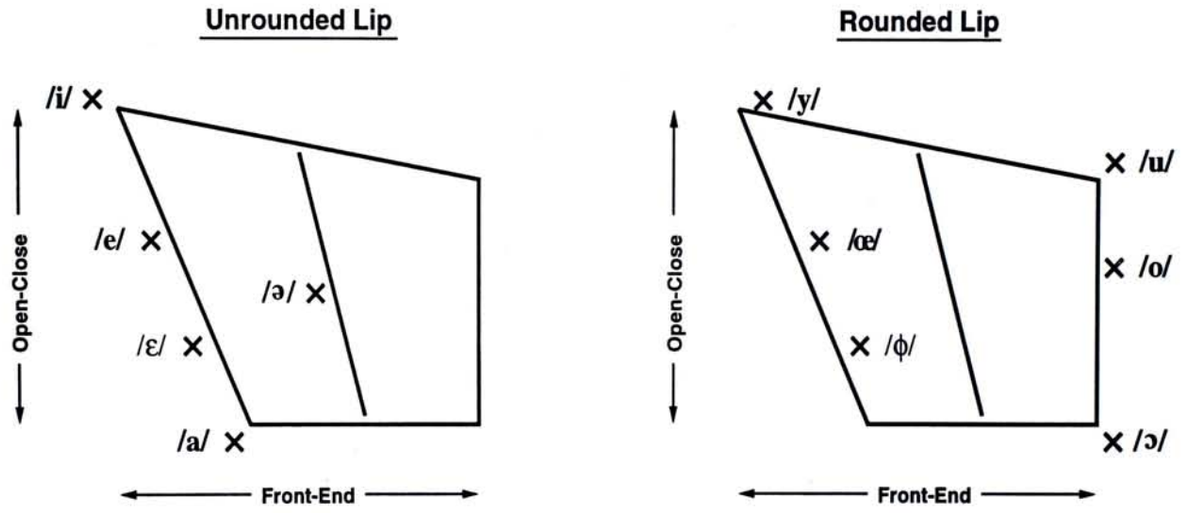

Figure 4.8: The coverage of the Cantonese vowels and diphthong targets within the cardinal vowel diagrams.

4.4 Summary

In this chapter, the neural network speech synthesis approach for the phone-based speech synthesis is elaborated. Basic functionality of this method is described.

The adoption of neural network in this synthesis method aims at the phone-based synthesis through the mapping capability. It can also provide non-linear approximation for spectral properties of signals in the transitions regions between templates. This approach also allows implicit, non-parametric storage of the speech templates. In addition, flexibility in the architecture of the synthesizer network enables the selection of input control parameters according to the specific requirements of the target system. The choice of articulatory control parameters provides additional flexibility and controllability in the synthesis process. Some important features and capability of the synthesizer network, such as voice correspondence, also depend on this specific type of control space.

In order to verify the capability of the proposed method, prototype networks of various structure have been constructed. It is found to be capable of synthesizing speech in a fairly simple way. Furthermore, subjective listening test has been carried out to assess the perceived quality of output. Detail of the test will be described in the next chapter.

Chapter 5

Prototype Implementation of the Synthesizer Network

In this chapter, a prototype system for demonstrating the functionality and potential of the synthesizer network will be described. Detail of the architecture and some fine-tune enhancement will also be discussed. Moreover, an informal subjective listening test has been conducted and the result will be presented.

5.1 Implementation of the Synthesizer Network

In order to verify the capability of the approximation network in speech synthesis, simple network models to imitate the vocal tract have been built. These network models are used in a prototype Cantonese synthesizer which serves as a vehicle to demonstrate the operation of the synthesizer network. Several networks of different types and architectures have been trained to examine the performance of the proposed synthesis technique.

In figure 5.1, the functional block diagram of the prototype synthesis system is shown. The synthesizer system basically accepts phonemic transcription and returns the corresponding speech signal.

Based on the phonemic transcription, control parameters targets (articulatory control, segmental duration and intensity) are retrieved from a look-up table. On the other hand, the pitch targets are retrived from another pitch table based on the specified lexical tones. These targets are then interpolated to form desired sequences of control parameters (articulatory parameter, pitch and amplitude). The articulatory control parameters

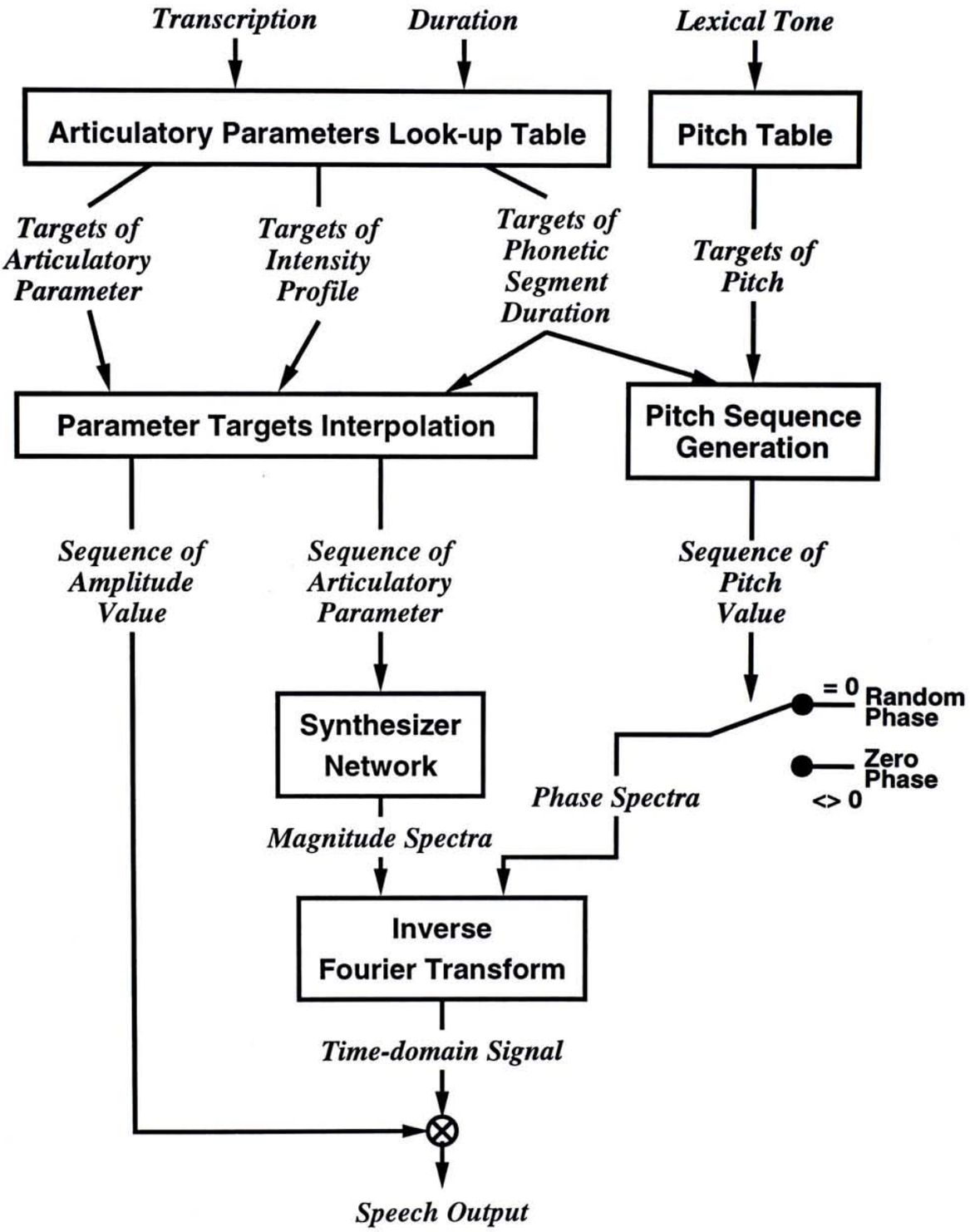

Figure 5.1: The block diagram illustrates the operation of the synthesis process, from transcription, duration and lexical tone down to the output of the final speech signal.

are fed to the synthesizer network with the magnitude spectra returned. The pitch sequence, on the other hand, controls the phase spectra and the pitch of the time-domain signal. The magnitude and phase spectra are nverse Fourier transformed to time-domain signals. These time-domain signals are then modified by the required amplitude profile and concatenated to form the final speech output.

5.1.1 Network Architectures

Two types of network are used to construct of the synthesizer network — feed forward backpropagation network and the Gaussian radial basis function network. They are used as the basic architecture for this multidimensional approximation problem.

Feed Forward Backpropagation Network Training

The first type of approximation network being investigated is the multi-layer feed forward backpropagation network. In order to obtain more approximation capability and to incorporate certain degree of non-linearity into the synthesizer network, sigmoid function is used as the activation function for the non-linear layers.

$$f(x) = \frac{1}{1 + e^{-k \cdot x}} \tag{5.1}$$

Equation 5.1 gives a bounded continuously differentiable function. It enables error backpropagation training to be used for the synthesizer network.

All the architectures under investigation have one input layer, two non-linear hidden layers with sigmoidal activation function and an output layer with linear activation function. The hidden layers make up the non-linear approximation for the synthesizer network. Whereas the linear output layer is essentially a linear combiner of the hidden layer outputs. It is used to produce unbounded final output from the synthesizer network

.

In many artificial neural network application, the least sum square error is usually employed as the optimization goal. Its advantage is that the problem is made simpler and a small difference between the target and estimation is guaranteed in the optimal sense. However, in speech processing, optimization of spectral information on the square

error does not guarantee good perceptual quality. A more commonly used error criterion is the Itakura-Saito distortion (ISD) measure which is defined as

$$ISD = e^V - V - 1 \tag{5.2}$$

In equation 5.2, V is the difference between the log-spectrum of the targeted and estimated signals. This distortion measure is essentially a dynamically weighting function that emphasizes the spectral region with larger magnitude. Whilst, a sum square error is a equally weighted difference criterion for distortion measure. In figure 5.2, the two errors are compared. It can be observed that the Itakura-Saito distortion is not symmetric about zero as the square error.

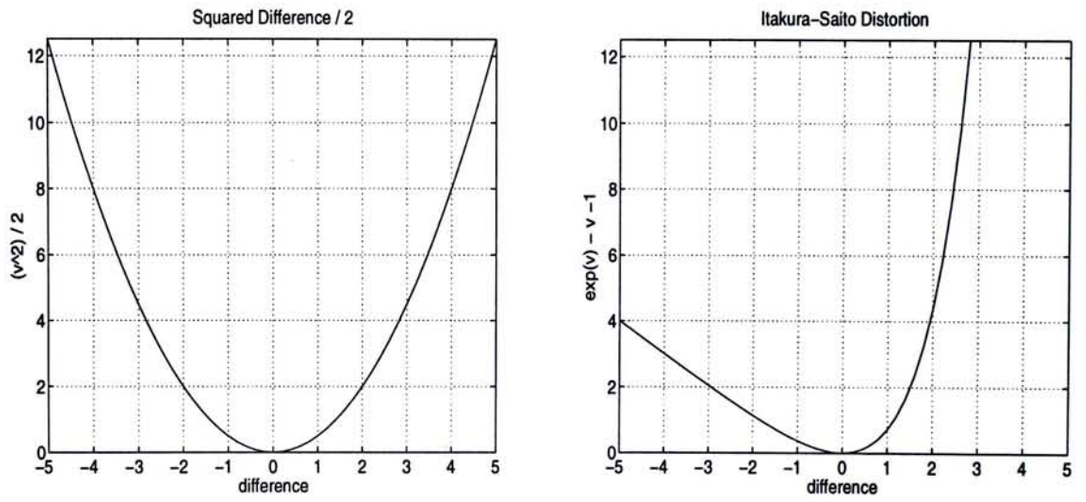

Figure 5.2: Comparison of the sum square error and Itakura-Saito distortion. Itakura-Saito distortion is dynamically weighted but it is not symmetric.

Modified Itakura-Saito Error Criteria Although the Itakura-Saito distortion measure is proven to be useful in speech processing, it cannot be directly applied as error function in backpropagation neural network training. This is because of the exponentially increasing slope of the error criterion makes the backpropagation training difficult to converge. [Appendix B.1, equationB.10]. Nearly all the emphasis is put on the regions with large target making the over-corrected regions difficult to be brought back to the desired solution. In order to obtain a more appropriate error function for use in artificial neural network, the ISD is modified. The modified Itakura-Saito Distortion (MISD) is

constructed with reference to the traditional ISD. It is symmetric about zero and therefore the weighting is the same for either target based (target – estimation) or estimation based (estimation – target) error estimation.

For a target spectrum t and an estimated spectrum a, we eliminate the constant term 1 and put balanced weights on both large target and estimation. As a result, we obtain

$$MISD = \frac{t}{a} - \frac{a}{t} + \ln \frac{t}{a} \tag{5.3}$$

Using log-spectrum difference, $V = \log t - \log a$, equation 5.3 becomes

$$MISD = e^{V} - e^{-V} + V \tag{5.4}$$

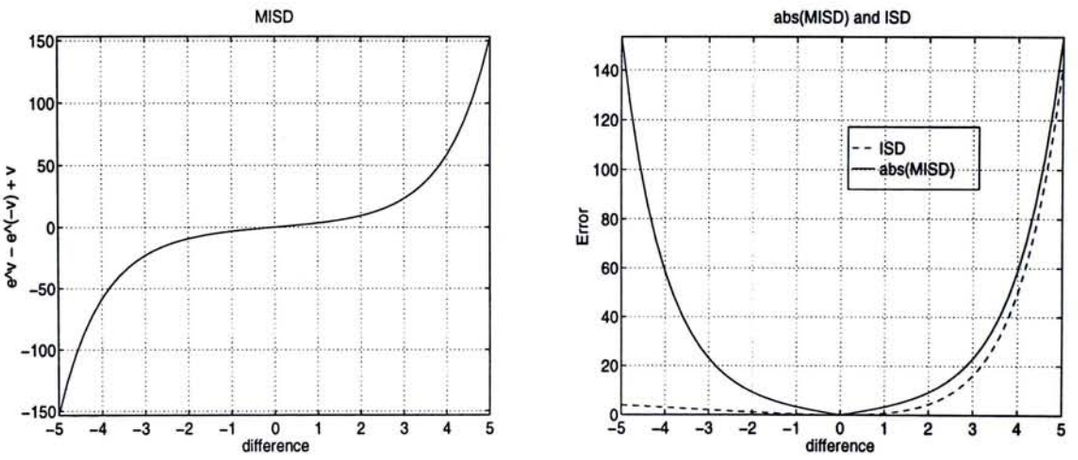

Figure 5.3: The modified Itakura-Saito Distortion and the comparison between MISD magnitude with the original ISD.

In figure 5.3, we can see that the MISD error function is an odd function about the origin. To ensure a magnitude matching in the spectra (similar to absolute error instead of linear error), the MISD used in the synthesizer network is therefore squared (absolute value is not chosen as it has differential discontinuity). Furthermore, it can also be shown that the gradient of the squared MISD is a symmetric function [Appendix B.2, equation B.16] which thus eliminates the trend of under estimation.

Another reason for using the MISD is that it retains the dynamic weighing property as in traditional ISD. While the ISD distortion generates estimation error by putting additional weighting to those regions with larger target value, the MISD effectively give

emphasis both on the spectral peaks in targets and false peaks in the estimation. Therefore, the estimated spectrum is expected to track the targeted spectrum much more reliably.

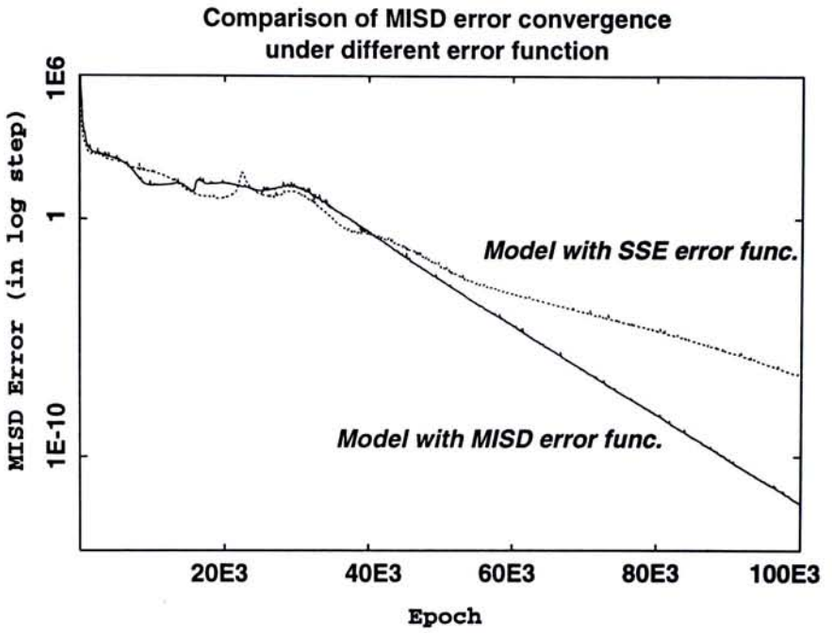

Figure 5.4: The error convergence of the backpropagation training for two networks. One uses the SSE error criterion and the other employs the MISD error criterion. The comparison shows that the SSE optimized network may show a slower convergence under the MISD error criterion.

Figure 5.4 shows the error convergence for the training of a network with SSE and MISD error function respectively. The comparison is made using the MISD distortion measure. The result indicates that the network with SSE error function converges more slowly when compared to that of the MISD. It is trivial because the SSE network is instructed to travel in a direction that has smaller SSE. Although the two error functions, have the same ultimate solution for null error, the difference in the error measure causes backpropagation training to travel with different paths. From the speech processing point of view, a more perceptual oriented error is desirable in order to obtain better quality.

It is found empirically that for a training set with 10 to 20 spectral templates at 8kHz sampling rate, a network size of around 3-10-20-32 to 3-20-40-32 is sufficient. As the amount of information increases with higher sampling rate, the network size is expected

to be increased. If more training templates are available, a larger network should also be used. This will enable the templates to be retrieved accurately upon presenting the corresponding input control during synthesis. However, it is believed that the required network size will soon grow slower as the redundancy in the templates increases. On the other hand, since backpropagation training can be trapped by local minima in the highly non-linear optimization problem, the network size should not be excessively large. Otherwise, it will be difficult to obtain desirable solution and the generalization property may be lost as well.

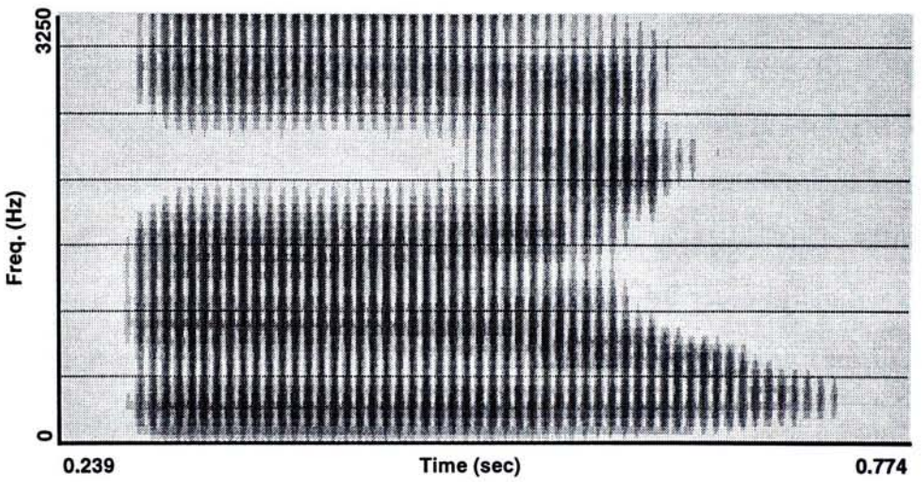

Figure 5.5: Example spectrogram of speech output (/nuai/) from the FFBP synthesizer network.

Error Backpropagation Training

In this synthesizer architecture, the multi-layer feed forward synthesizer networks are trained using the traditional error backpropagation technique. Both sum squared error (SSE) and modified Itakura-Saito distortion (MISD) have been used as the network error function. Networks of different size have also been trained using the training templates. In general, the resultant synthesizer shows that trained synthesizers based on the MISD possess better output quality for similar errors.

Radial Basis Function Network Training

Radial basis function is a class of basis functions that is popularly used in the function approximation problem. It is applied to the synthesizer network with an aim to investigate its capability in the approximation and interpolation of speech spectral templates.

In this prototype, Gaussian radial basis function (GRBF) approximation network is used. It gives the optimized result of approximation networks under the sum square error criterion with regularization on the second derivative of the basis function. The resultant networks are found to be capable of synthesizing speech signal with good approximation.

The GRBF network employed is constructed using a multi-layer architecture. It consists of an input layer, a hidden layer of radial basis functions and an output layer of neurons that are linear combiners. In general, the optimization process in the training stage is to find the optimal basis centers and weights. However, simultaneous optimization of the centers and weights is a highly non-linear and complex task. Therefore, it is common to choose centers from the training templates and then optimize the weights. This simplification relies on the assumption that among the training template there are representative templates for the targeted space of approximation.

Orthogonal Least Square Center Selection There are many different methods for the selection of basis centers given a set of training templates. Some of these are random based while others may make use of pruning and growing technique among the candidate centers. In the radial basis function synthesizer network, an orthogonal least square center selection scheme is adopted [104]. It selects the appropriate centers among the training data on individual basis by minimizing the least square error of the network output.

The orthogonal least squares (OLS) learning algorithm is given by

$$d(t) = \sum_{i=1}^{M} p_i(t)\theta_i + \epsilon(t) \tag{5.5}$$

where d is the desired output, p_i are the basis which are function of the network input x_i, θ_i are the weighting parameters and ϵ is the noise. The advantage of OLS is that it can assist the selection of basis center through the classical Gram-Schmidt procedure [104].

An error goal or maximum number of basis can be specified as the stopping criterion. More details are provided in Appendix C.1.

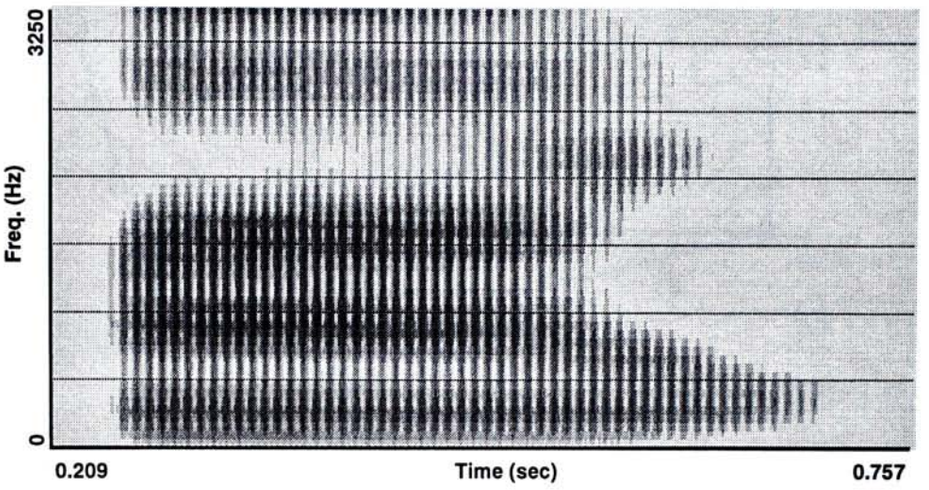

Figure 5.6: Example spectrogram of speech output (/nuai/) from the RBF synthesizer network.

Besides the basis center, the spread of the basis function can also be changed. The spread of the basis is the variance, σ^2, in the Gaussian function. It determines the degree of smoothness assumption in the approximation. In the OLS algorithm, the orthogonal selection is based on the specified spread, while the choice of proper spread is based on the transition quality of the resultant network and is done empirically. Figure 5.6 shows an example of spectrogram from RBF network output.

5.1.2 Spectral Templates for Training

Spectral data of phone units from recorded speech are used as training templates to train the FFBP and RBF synthesizer networks. The articulatory control parameters are estimated according to the location of the phone units in the combined space of the vowel quadrilateral.

Template Selection

In general, for a particular target language (Cantonese in this example), the vowels near the cardinal vowels in the vowel diagrams should be included. For example, the phone units of the Cantonese vowels are all included in the training data [Figure 4.8].

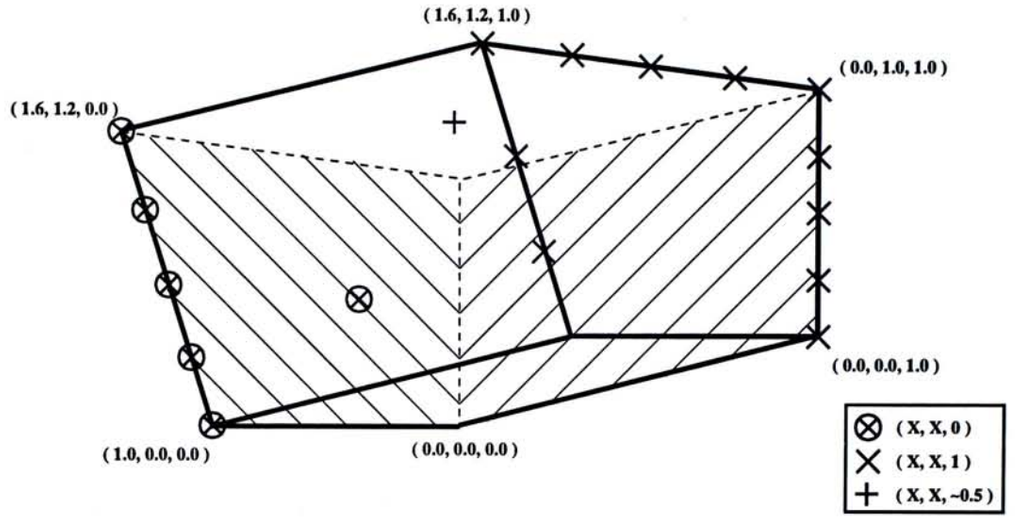

Figure 5.7: The coverage of an examples training set (18 training templates) for Cantonese in the 3-dimensional articulatory space is shown. All of the simple vowels are included with certain amount of phone units extracted from diphthongs.

Description	Articulatory Parameters
/a/	(1.0, 0.0, 0.0)
between /a/ & /i/	(1.15, 0.3, 0.0)
between /a/ & /i/	(1.3, 0.6, 0.0)
between /a/ & /i/	(1.45, 0.9, 0.0)
/i/	(1.6, 1.2, 0.0)
between /u/ & /i/	(0.8, 1.1, 0.6)
between /u/ & /y/	(0.4, 1.05, 1.0)
between /u/ & /y/	(0.8, 1.1, 1.0)
between /u/ & /y/	(1.2, 1.15, 1.0)
/j/	(1.2, 0.4, 1.0)
between /j/ & /y/	(1.4, 0.8, 1.0)
/y/	(1.6, 1.2, 1.0)
/c/	(0.0, 0.0, 1.0)
between /c/ & /u/	(0.0, 0.25, 1.0)
between /c/ & /u/	(0.0, 0.5, 1.0)
between /c/ & /u/	(0.0, 0.75, 1.0)
/u/	(0.0, 1.0, 1.0)
/r/	(0.5, 0.5, 0.0)

Table 5.1: The table lists out the phone units used in training the prototype synthesizer network together with their articulatory control parameters.

In order to assist the synthesizer network to establish the unique relationship between the input control and spectral output, other pertinent information is desirable. Phone units extracted from diphthongs will provide valuable details about the properties of the transition regions [Figure 5.7]. Theoretically, the more the templates used, the better the approximation. However, the network size and the training resource requirement will also increase with the training data. The phone templates used for training the prototype networks are listed in table 5.1 for reference.

Template Pre-processing

Speech data used for training synthesizer network are pre-processed at the front end. Phone units extracted from recorded speech are converted to spectral domain representation by Fourier transform. The spectra are then interpolated to twice the length of the targeted network output (c.f. magnitude spectrum of real signal is an even function). The first half of the magnitude spectrum is used as the spectral templates, Z_i, in the training templates. Together with the estimated articulatory control parameters, X_i, the spectral templates form the training templates pairs, $\{X_i, Z_i\}$, for the synthesizer network training. The pre-processing applied to extract phone units is shown in figure 5.8.

5.1.3 System requirement

In view of the memory and computation, the proposed synthesis method places moderate requirement on the resources. This method takes only little storage for the speech templates by using phone units. Further reduction is achievable by taking advantage of the redundancy in the training templates. The largest requirement on memory comes from the table for transcription to control parameters conversions. It stores up the target points for the articulatory control parameter profiles (Depending on the syllables, an average of ten set of articulatory control parameters targets are needed for each syllable.).

On the other hand, the computation complexity for this synthesis method is mild. The mapping of control parameters to output spectrum is simply a non-linear mapping by trained FFBP neural network or RBF network. The most intensive tasks, however, are the interpolation of the parameters from the targets to desired length. The spline

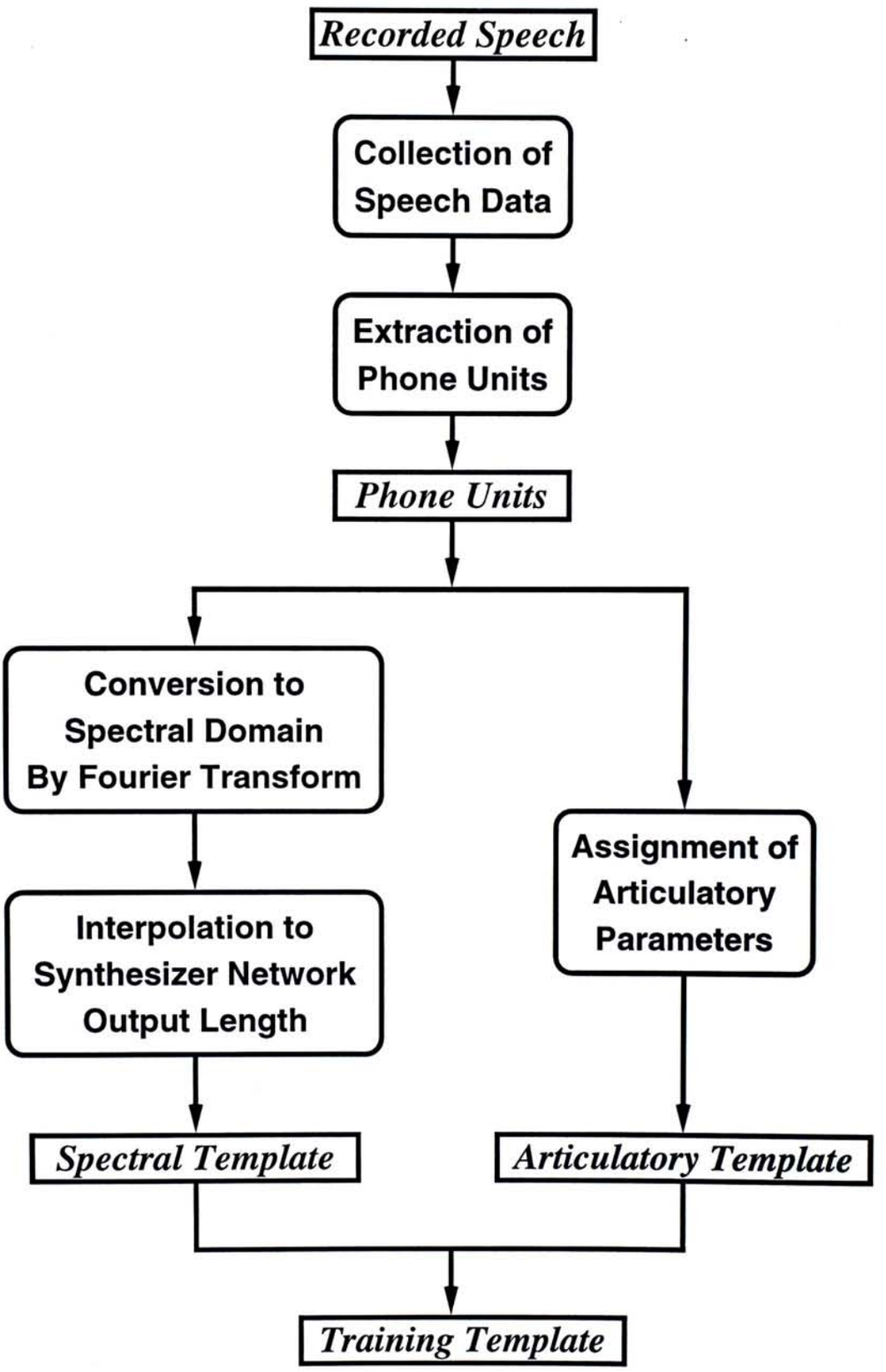

Figure 5.8: The pre-processing work applied to the phone units before using in the network training process.

interpolation gives a smooth parameters profile at the cost of more computation works. The computation load from the inverse Fourier transformation of the magnitude spectrum to speech signal can be relaxed by using fast transformation algorithms.

In this prototype system, aiming at proofing the feasibiliy of the method, the complete system is implemented on the Matlab [1] environment. From the stage of extraction of phone templates from recorded speech, tranining of the synthesizer network, and the final synthesis process, it is performed using Matlab (With an exception that the training of the backpropagation network is performed using C programs for reducing the training time.).

In the synthesis stage, the complete process is performed using the Matlab environment. The control parameter targets are stored in M-files as variables to take the role of a real look-up table. In addition, the interpolation of parameters, the neural network mapping and inverse Fourier transform are all carried out using Matlab. In general, from the stage of giving the transcription up to obtaining the output speech file, it takes around one second for 500 ms of output speech on a Sparc 20 [2] platform.

The slow synthesizing speed is believed to be caused by the slow looping performance of Matlab. Time ticking of major steps of the synthesis process also shows that the interpolation process for articulatory control parameters take more than half of the total processing time. To improve the processing speed of the system, the synthesis should be implemented using programming language such as C. This can certainly improve the speed significantly because the looping can be inplemented with integer instead of floating point [3]. This will reduce the time spent on the spline interpolation and resizing the spectrum length of the syntehsizer network. To further enhance the performance, the synthesis process can be implemented using hardware. There are places that parallel processing is feasible and significantly improve the overall speed performance. The interpolation, neural network mapping and the inverse Fourier transform are already capable of fast parallel hardware implementation and can ensuring better speed performance.

[1] Matlab is trademark of Mathworks
[2] Sparc 20 is a trademark of Sun Microsystem
[3] Integers in Matlab are implemented through flint – floating point integer.

5.2 Subjective Listening Test

Subjective listening test has been conducted to assess the quality of the synthetic speech. The test concentrates on the acoustic (segmental) properties rather than the prosodic (suprasegmental) properties of speech signal. The aim is to grade the naturalness of the output speech signal from the synthesizer network of different architectures. The listening test is based on that described in [47]. However, cardinal grade point is used instead of nominal grade, and listeners are requested to suggest subjective grades for the presented speech signal.

5.2.1 Sample Selection

In order to eliminate any suprasegmental features affecting the assessment of the listeners, only short speech segments are used in the test. The test data are divided into three ***groups*** [Figure 5.9]. Each group has various ***set*** of data and each set has samples of different ***types***. The three groups of speech data are **vowel**, **consonant-vowel (CV)**, and **diphthong**.

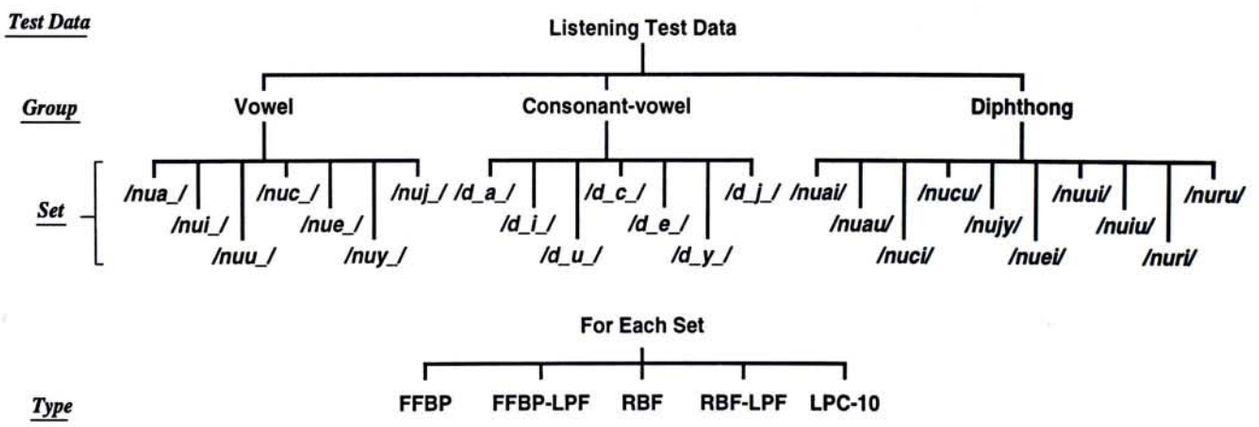

Figure 5.9: The hierarchy shows the test data included in the test and also the naming convention of the data in the hierarchy.

Vowel In the vowel group, all of the seven simple vowels in Cantonese are included [Table 5.2]. The vowels are set to the lower-going tone, tone 6. Since the synthesizer network generates these vowels basically by templates retrieval, the basic capability of the

synthesizer network in speech signal synthesis is demonstrated. The seven vowel samples included in the test are : ***/nua_/***, ***/nui_/***, ***/nuu_/***, ***/nuc_/***, ***/nue_/***, ***/nuj_/***, and ***/nuy_/***.

Consonant-vowel syllable The group of consonant-vowel (CV) syllables encompass all seven Cantonese simple vowel finals. They all start with the initial consonant /d_/ with the upper-going tone, tone 3. This group aims at testing the perceptual quality of the synthesized signal in CV syllables. Different sets of CV syllables are ***/d_a_/***, ***/d_i_/***, ***/d_u_/***, ***/d_c_/***, ***/d_e_/***, ***/d_j_/***, and ***/d_y_/*** [4]

Diphthong The diphthong group includes all ten Cantonese diphthongs. They are intended to investigate the vowel-to-vowel transitions in compound vowels of Cantonese syllables. The lower-going tone, tone 6, is used in the diphthongs. The ten sets of speech data are ***/nuai/***, ***/nuau/***, ***/nuci/***, ***/nucu/***, ***/nujy/***, ***/nuei/***, ***/nuui/***, ***/nuiu/***, ***/nuri/***, and ***/nuru/***.

Groups								
Vowels			CV syllables			Diphthongs		
Set	**IPA**	**Tone**	**Set**	**IPA**	**Tone**	**Set**	**IPA**	**Tone**
/nua_/	/a/	6	/d_a_/	/da/	3	/nuai/	/ai/	6
/nui_/	/i/	6	/d_i_/	/di/	3	/nuau/	/au/	6
/nuu_/	/u/	6	/d_u_/	/du/	3	/nuci/	/ɔi/	6
/nuc_/	/ɔ/	6	/d_c_/	/dɔ/	3	/nucu/	/ou/	6
/nue_/	/ɛ/	6	/d_e_/	/dɛ/	3	/nujy/	/ɸy/	6
/nuj_/	/œ/	6	/d_j_/	/dœ/	3	/nuei/	/ei/	6
/nuy_/	/y/	6	/d_y_/	/dy/	3	/nuui/	/ui/	6
						/nuiu/	/iu/	6
						/nuri/	/ɐi/	6
						/nuru/	/ɐu/	6

Table 5.2: The speech samples used in the listening test includes vowels, CV syllables and diphthongs.

[4] /d_i_/, /d_u_/ and /d_j_/ does not occur in formal Cantonese (except colloquial phrases). They included only to test the CV synthesizing ability of the synthesizer network.

Types of speech samples In each set of syllables, there are five different types of data. Some of these are output signal used for testing (RBF, FFBP) while others are used to give reference grading level (LPC-10,XXX-LPF) for reference. The five different types of speech outputs are

1. feed forward backpropagation synthesizer network output **(FFBP)**,
2. low pass filtered feed forward backpropagation synthesizer network output **(FFBP-LPF)**,
3. radial basis function synthesizer network output **(RBF)**,
4. low pass filtered radial basis function synthesizer network output **(RBF-LPF)**,
5. LPC-10 coded real speech **(LPC-10)**.

As a result, there are 3 groups with different number of speech data sets (7, 7, and 10) and there are 5 different types in each set that makes up a total of 120 [5] segments of speech signals. During the test, all these segments are presented to the listeners for assessment.

5.2.2 Test Procedure

There are totally five listeners in the test and none of them are involved in speech processing research. Speech experts are *intentionally excluded* such that the result of the listening test will be more reliable. People with experience in speech processing may have bias in the assessment depending on their particular training.

For each listener, the speech signals are presented through a headphone and are requested to grade them one after the other. They are allowed to replay particular sample or repeat the whole set of samples to aid them in proper assessment.

The presentation order of the types of outputs within each set is initially randomized. The resultant order will be used as the presentation order to all listeners. This randomization can reduce the chances that the listeners perceive regular pattern in the presented speech across different set that may affect their grading. However, the presenting orders

[5] 3 x (7 + 7 + 10) x 5 = 120 speech segments

for different listeners are the same so that listeners can compare the different type of speech more easily.

After initialization, listeners are instructed to assess the speech output based on their subjective feeling. A guideline scale is given to the listener as shown in table 5.3.

0 - 2	**2 - 4**	**4 - 6**	**6 - 8**	**8 - 10**
Unacceptable	Intelligible	Fair	Comfortable	Perfectly human

Table 5.3: The guideline scale for the listener in the informal subjective listening test.

The output presented to the listeners are either digitally processed speech signal or synthesized output in digital form. The specification of the digital speech and the hardware used in the test are given as follows :

- 8000 kHz sampling rate with 16 bit signed integer precision,
- SparcUltra Build-in Audio device, and
- Sony MDR-V200 Stereo Headphone

5.2.3 Result

The results of the subjective listening test are summarized in this subsection. Table 5.4 to 5.6 and figure 5.10 to 5.12 give the results for the vowel test, consonant-vowel test and diphthong test respectively.

	/nua_/	*/nui_/*	*/nuu_/*	*/nuc_/*	*/nue_/*	*/nuj_/*	*/nuy_/*
FFBP	5.5	3.75	2.7	5	4.25	4.275	3.5
RBF	5	3.5	3.5	4.875	4.625	4	3.25
FFBP-LPF	5.05	2.75	2.95	4.875	4.875	4	4.375
RBF-LPF	5.075	3.625	3.25	4.75	4.5	4.2	4
LPC-10	2.975	2.425	1.75	1.7	2	1.525	1.375

Table 5.4: Listening test result of vowel test

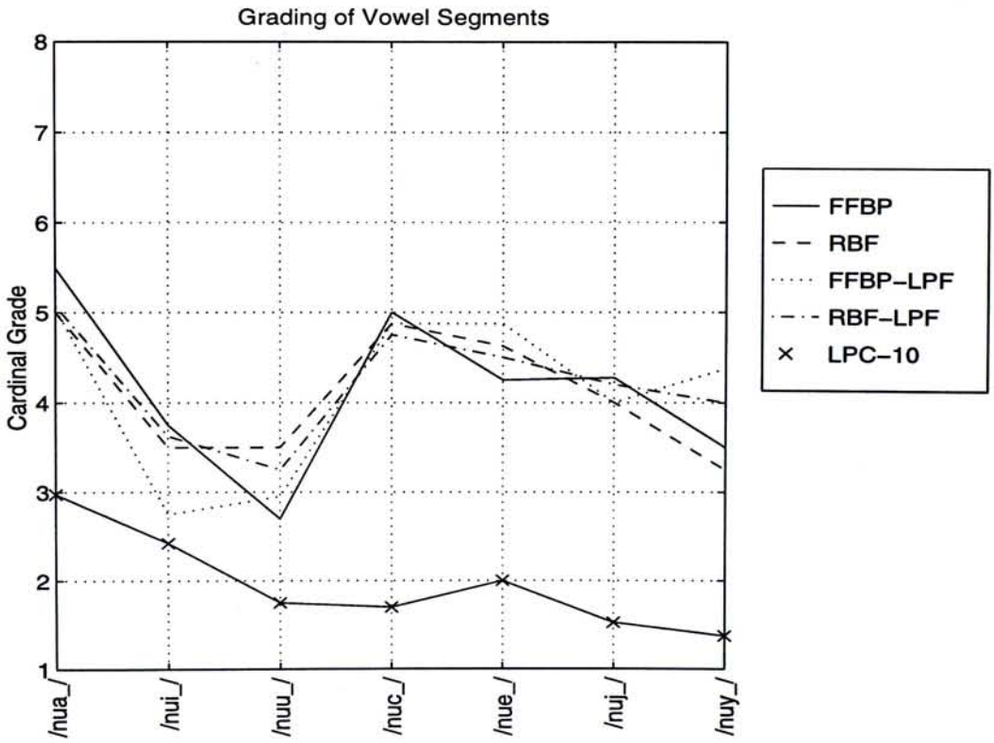

Figure 5.10: Listening test result of vowel test

	/d_a_/	*/d_i_/*	*/d_u_/*	*/d_c_/*	*/d_e_/*	*/d_j_/*	*/d_y_/*
FFBP	5.175	4	4	4	4.475	3.225	3.975
RBF	4.75	4	2.875	3.375	3.975	2.975	4.125
FFBP-LPF	5.375	2.75	2.875	3.375	3.225	2.725	3.125
RBF-LPF	4.75	3.5	2.875	3.5	4.225	3.725	3.5
LPC-10	2.125	2.625	1.25	2	1.875	1.25	2.125

Table 5.5: Listening test result of consonant-vowel test

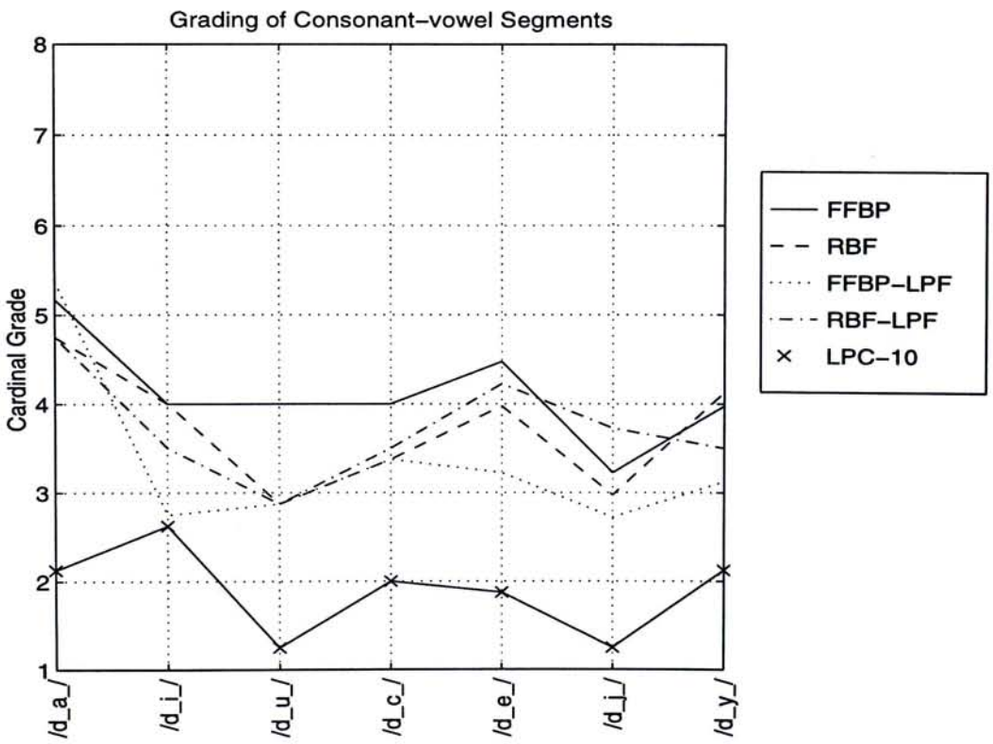

Figure 5.11: Listening test result of consonant-vowel test

	FFBP	RBF	FFBP-LPF	RBF-LPF	LPC-10
/nuai/	7	6.875	7.875	7.625	4.5
/nuau/	5.25	5	5.125	5.25	3.5
/nuci/	6.25	5.25	5.75	4.875	4
/nucu/	4.625	5.375	4.5	4.5	3.75
/nujy/	5.75	4.5	5.25	4.875	4
/nuei/	5.875	5.75	5.5	4	4.625
/nuui/	4.875	5	5.5	4.75	3
/nuiu/	4.625	5.125	4.75	4.875	3.75
/nuri/	6.25	5.5	6.75	6.5	4.5
/nuru/	5.5	5.875	5.125	5.875	3.125

Table 5.6: Listening test result of diphthong test

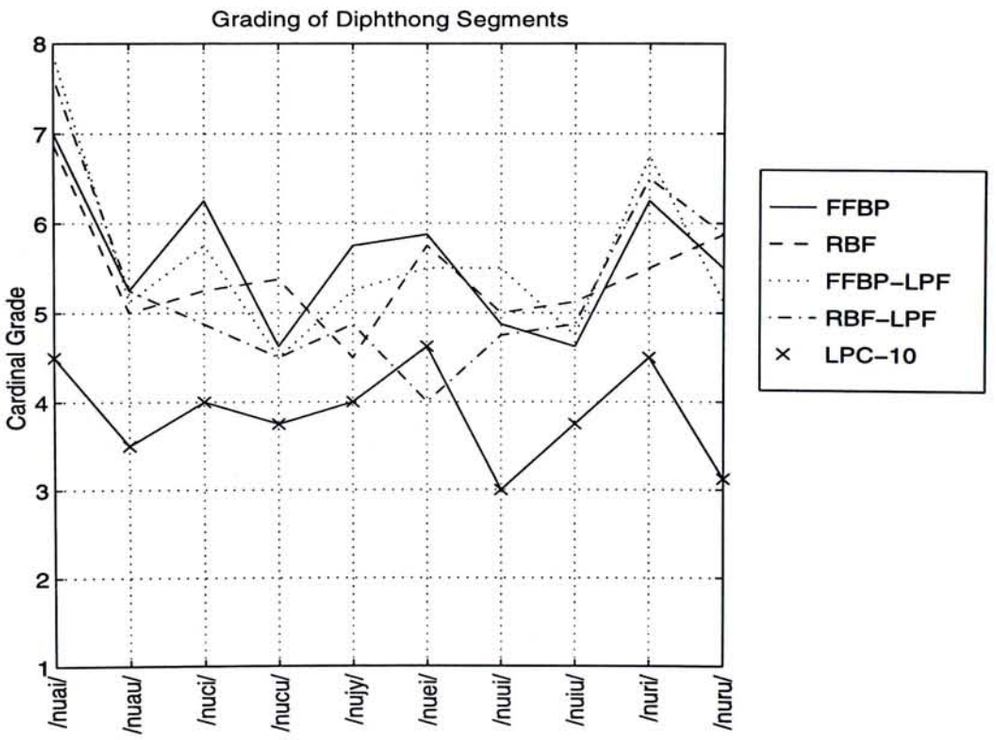

Figure 5.12: Listening test result of diphthong test

5.2.4 Analysis

From the listening test, it is observed that the speech outputs of the synthesizer networks are in general of fairly good quality. The assessment is based on subjective listening of the speech quality. Although it is only a small scale test, the results indicate certain degree of consistency. As expected, the grading for synthesizer network outputs increases as the the segment of speech gets larger and reach fairly high level for some of the diphthong units. Comment on individual sample set will be made in the following paragraphs.

Vowel

The result of the vowel test shows that the grades of the synthesized output from different networks are more or less the same. Although they give consistent superiority over the LPC-10 coded speech, they cannot be regarded as possessing high naturalness. The lower grades in single vowel segments are expected because the acoustic properties of short segments lack variations. This makes them sound monotonous to any listener even the coded speech. Although the LPC-10 coded speech is used as reference level only, this test indicates that the synthesizer networks are capable of producing output speech with quality better than the commonly used LPC-10 low bit-rate coded speech. On the other hand, the decline in quality for the */nui_/*, */nuy_/* and */nuu_/* indicates that isolated close vowels sound more unnatural than others.

Consonant-vowel

For the consonant-vowel syllables, the listening test result is more or less the same as that of the vowel test. This is because these are still short segments that lack acoustic variations. However, the absence of degradation for these CV syllables demonstrate the capability of the synthesizer network in CV syllables synthesis. To certain degree, it also shows that the network approach is capable of synthesizing perceptually acceptable transitions in the CV transitions.

It should be reminded that in this CV syllable test, only the alveolar initial /d_/ is tested. This is only selected as an example for illustrating the perceptual quality of syllables with CV transitions synthesized by the network approach. In synthesizing this

type of CV transition, articulatory correspondence is used for the transition from initial. Details will be given in the next chapter on articulatory control of the speech synthesis problem.

Diphthong

For the diphthong test, the result agrees well with theoretical prediction. The subjective quality grading increases a great deal over the short, near stationary speech utterance – simple vowel, CV syllable. The inherent acoustic variations in the diphthong units seem to give the feeling of better perceptual quality. Figure 5.12 shows that there is a significant increase in grading for both the synthesized speech and the coded speech.

The simultaneous improvement in the perceptual quality of both types of speech output indicates that the synthesizer networks have successfully synthesized utterance of perceptually acceptable transitions between the vowel to vowel transitions in diphthongs. The result also shows that the quality of the diphthongs near the region with more training templates [refer to figure 5.7] is better. This is expected and verifies the theoretical prediction that more training templates will generally give better quality. Moreover, the test result clearly shows that the subjective grading for the network synthesized diphthongs is quite high.

The result shows that this approach can synthesize fair quality short segment speech signal with a small size synthesizer network and only small amount of templates. This indicates that this approach has successfully generated the spectral features of the required speech segments from information in the training templates. Moreover, the test also strengthens the believe that with appropriate property variations, longer speech segments sound more natural.

5.3 Summary

In this chapter, prototype implementation of the synthesizer network and its training has been described. Moreover, a performance test for the proposed network approach for phone-based speech synthesis was described. The capability of this method is testified through the prototype implementation and subjectively assessed through an informal listening test. The prototype implementation aims at demonstrating the feasibility of acoustic speech signal synthesis through neural network.

The result of the informal subjective listening test shows that fair to comfortable speech quality can be achieved. The outcomes also agree with predictions made based on theoretical reasoning.

In the following chapter, some other implementation issues and simplifications based on the articulatory control will be described. It will make better use of the articulatory control for the synthesis of speech signal.

Chapter 6

Simplified Articulatory Control for the Synthesizer Network

Sufficient external control needs to be provided to a synthesizer network in order that it can produce arbitrary segments of speech with high quality. Although a number of control techniques for speech synthesis have been proposed and investigated, the most natural and intuitive choice is the articulatory-based control method. Articulatory control not only provides effective control that simulates the human speech production mechanism, but also enhances the flexibility and controllability of speech synthesis. It can be used to model different types of variations in the actual speech production process. This makes it particularly attractive to be developed as the ultimate solution to produce natural human speech.

Figure 6.1 illustrates the general idea on the synthesis control – the articulatory control for the spectral features and some prosodic features control.

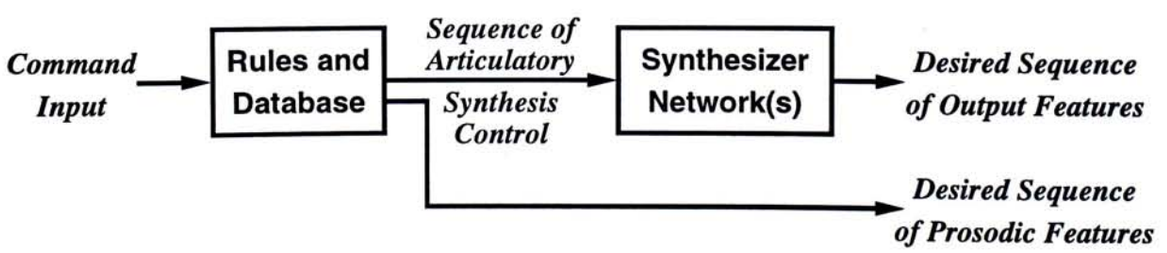

Figure 6.1: The use of articulatory control parameter for the synthesis control in neural network speech synthesizer. The control sequences to the synthesizer network(s) and post-processing parts are generated according to input command and based on pre-stored information.

In this chapter, we shall discuss some articulatory control schemes and propose a simple way of using the articulatory parameters derived from the combined vowel spaces as synthesis control for the NN synthesizer network. Some practical issues on the implementation of this method will also be elaborated.

6.1 Coarticulatory Effect in Speech Production

In speech production, acoustic signals are generated by the articulators modulated phonation. However, for various reasons, there are differences in the signal properties between isolated utterance of syllables and continuous speech. These include both the **acoustic** and **prosodic** properties [35, 37]. The causes of these variations are mainly physical limitation of articulators as well as psychological effects. The contextual relationship of the utterances can create mental linkage that will make the speaker subconsciously carry some properties of one acoustic segment over to other "related" segments.

6.1.1 Acoustic Effect

The acoustic coarticulatory effect refers to those acoustic variations of speech segments due to other speech segments. In continuous speech, there are usually changes in between the uttered segments and some of the acoustic properties of one segment may be carried to other segments. These variations in properties are mostly linguistically related.

The causes of the coarticulatory effect in continuous speech are multi-fold. First of all, for the speech segments that are physically adjacent in time, the reason of coarticulation is obvious. Due to the difference in place and manner of articulation, there must be a finite duration for the transition. In general, the faster is the speaking rate, the greater the coarticulatory effect. With larger difference in the place and manner of articulation, the effect will be more prominent. These changes are due to the physical limitation of the speech production system.

On the other hand, there are also psychological causes for the coarticulatory effect. Between adjacent syllables/words, people may try to make preparation for the next articulated word. This results in both forward and backward transfer of acoustic properties between segments. It is also possible that properties of all related segments are

modified. Acoustic property transfers are, therefore, not only common across syllables without inter-syllable pause, but also found in regions with inter-syllable pauses.

6.1.2 Prosodic Effect

The prosodic coarticulatory effect in continuous speech comprises of several different aspects. These include interference in the pitch profile and also the timing of segment units and pauses between syllables.

The pitch profile of syllables in a spoken utterance is determined by several factors : basic lexical tone of syllables, habitual changes of tones in phrases, intonation of the utterance, and coarticulatory variation of the pitch profile across syllables. The order of the factors also represent their importance in descending order.

The determination of timing property of segments and pauses in spoken language a fairly complicated issue. The basic phonemes, habit and intonation are crucial factors. In addition, it is also changed by coarticulation. Under different context, the segment and pause duration may change accordingly.

Similar to acoustic coarticulatory effect, the coarticulation of prosodic features are also brought about by physical and psychological causes. The finite transition time for the change of status of the vocal cord hinders any sudden change in pitch of vibration. Therefore, the starting and terminating pitch profiles of adjacent syllables may be interfered by any difference in lexical tones. Psychological effects may also be observed from the change of pitch profiles of syllables in different phrases, either in continuous speech or in isolation.

6.2 Control in various Synthesis Techniques

Synthesis controls in various speech production methods not only provide the necessary control for deriving the target speech signals, some of them also allow detailed model of speech variations due to coarticulation. Available control strategies for some common synthesis methods are given below.

6.2.1 Copy Concatenation

Copy concatenation is so far the simplest synthesis method. It also provides the most natural speech output for applications with predefined speech sounds. However, special methods could be employed to vary the acoustic and prosodic properties to simulate the coarticulatory effects.

To generate arbitrary speech output from stored templates, the synthesizer must put together pre-selected templates from the inventory. However, even for a moderate size inventory, the concatenated speech output will sound unnatural due to possible abrupt transition between templates. Simple solution to this problem is to apply interpolation to the templates. Improved result can also be achieved by including the transitions in inventory or even construct new template database based on the transitions (such as diphone, demi-syllable and triphone databases).

However, interpolation alone cannot solve the problem of required variations in speech segments. Change of phone unit properties due to coarticulatory or allophonic variations must be added to inventory if desired. Combining these solutions together, a large synthesis template database is inevitable. The storage and searching problems will hinder implementation in smaller systems.

In concatenation based synthesis, the change in pitch and timing of speech templates for target output are usually achieved by warping, insertion and deletion. For better result, Moulines et. al [4] have proposed the use of PSOLA for either time domain or frequency domain modification of speech on both time and pitch scale.

6.2.2 Formant Synthesis

In formant synthesizer, the acoustic variations of speech signals are usually achieved by the variation of the formants and bandwidths. Transitions are modeled by corresponding transition of formants and bandwidths.

The success of formant synthesizer model depends on the conformance of different control rules applied to model the actual transitions and variations. Simple rules for the formant and bandwidth transitions such as linear and sigmoidal transition paths are commonly used.

For the prosodic changes in pitch and timing, the change is simply achieved by varying the frequency and repetition of synthesizer excitation pulses. Formant synthesizer usually provides better control over the concatenation based approach. It also provides better models for the variations of speech signals which are prescribed by proper rules. The intelligibility of formant synthesizer are usually guaranteed by proper formant database of basic synthesis units such as phones or selected common allophones.

6.2.3 Articulatory synthesis

In articulatory synthesis, the acoustic variations of the speech output is achieved by controlling the area function or articulator positions. Since articulatory synthesis is trying to simulate the real speech production mechanism, proper control of the articulator positions can guarantee good transitions between segments. Moreover, allophonic variations can also be reliably simulated by corresponding variations in the articulator positions.

In general, segmental transitions and coarticulatory or allophonic variations of speech signal is achieved by controlling the articulatory parameters. The merits of this type of synthesis control are intuitive and highly controllable over simple concatenation.

6.3 Articulatory Control Model based on Vowel Quad

In the proposed synthesizer network, the input parameters of the network are articulatory based. These quasi-articulatory parameters represent a simplified articulatory model. It does not pose any limitation on further enhancement of the controlling model whenever deem necessary.

The articulatory model is derived from the vowel space of the vowel quadrilaterals [figure 6.2]. It is made up of three parameters that have been found to be sufficient for the generation of vowels in most languages. They include **front-end position**, **openness** and **lip-roundness**.

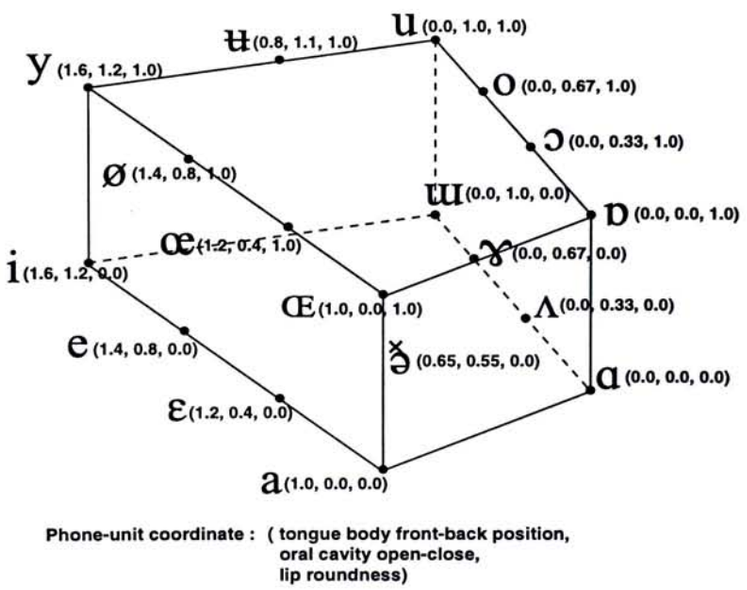

Figure 6.2: The vowel space of the combined vowel diagram is used as the control space for the simplified articulatory model. It provides an example articulatory control space for the synthesizer network.

With this articulatory control model, the complicated relationship between the articulatory space and spectral domain of phones can be efficiently modeled by providing training templates to the synthesizer network. It equips the synthesizer network with control flexibility and controllability similar to articulatory synthesis. In general, the trajectories of the articulatory parameters determine the transition of the spectral properties. It can be used for the modeling of coarticulatory and allophonic variations of speech signals through proper parameters control.

In addition, this control model effectively provides transition approximation for the phone-based copy concatenation synthesis. The reliable templates mapping through the synthesizer network makes it function like a phone-based concatenation synthesizer. The transition between the target segments are either obtained by concatenation of intermediate phone templates (if included in training data) or by approximating via the synthesizer network.

6.3.1 Modeling of Variations with the Articulatory Control Model

The simplified articulatory control parameter space is used for the modeling of spectral variations in speech. Although the reduced space provides far less controllability than the detailed area function, it provides a simple prototype example for illustrating the capability of this approach. As shown in figure 6.2, the simplified articulatory control is in a three dimensional space. Extension to a higher dimension hyper-space is feasible if that is essential to a specific applications.

From the trajectories of the required articulatory parameters, it is observed that the center of the vowel-space is not common in the trajectories in Cantonese speech utterance. Exception is found for the starting vowel /r/, which itself is located near the center. The related finals are */ru/*, */ri/*, */rn/*, */rt/*, ... Observation also shows that farther away from the center, the positions are more probable to lie within the articulatory parameters trajectory of continuous speech. Therefore, the trajectory of the control vector concentrates on the border of the vowel space except central vowels or short transition in nearby regions. This phenomenon can be explained by the fact that central voice is rare in Cantonese and there is much less training templates near the middle of the space.

Variation Modeling

Coarticulatory In this articulatory control, coarticulatory effect can be modeled in a similar manner as articulatory synthesis. It is practically a three parameter articulatory control for speech synthesis.

To simulate the acoustic coarticulatory effect in speech utterance, it may be done as interference between the adjacent segmental targets. These targets are retrieved from database for the specific speech segments. In continuous speech, transition of control vectors between the adjacent segmental targets is used to simulate the actual acoustic transition. Under extreme cases, the targets may be varied according to the other segmental target to reflect the actual coarticulatory effect in fluently uttered speech.

Allophonic : For the acoustic variations in allophones, the changes of targets are usually determined by some preset values based on the articulation properties of the related segments. Under the preset direction of variations for the segment targets, the synthesis control model will generate the desired acoustic variations through the synthesizer network.

Segmental transition : In vowel–vowel segmental transitions such as diphthongs, the synthesis control model can not only implement the allophonic variations of the vowel segments, but also realize the vowel to vowel transition in diphthongs. Based on the two preset targets in the diphthong and the control model, the synthesizer network provides approximation of the transition. In case the specified targets do not correspond to any training templates, the synthesizer network will attempt to produce such acoustic variations based on information derived from the training templates.

Similarly, consonant-vowel (CV) and vowel-consonant (VC) transitions can be generated. The production of consonants is facilitated by their voice correspondence together with prosodic control. The rules determining these correspondences and prosodic controls are already predefined in the database. Variations in control vector for allophones of consonants are stored as varied target vectors or correspondences [Section 6.4]. The actual acoustic variations are hence derived from the articulation positions of the consonant allophones.

6.4 Voice Correspondence :

In order to take advantage of the articulation process and to provide further simplification in the system implementation, ***voice correspondences*** are employed. The voice correspondence technique is based on the **place of articulation** of the original phone units and that of the referencing phones. *Voice correspondence is the assignment of articulatory parameters to some speech sounds that cannot be exactly represented by the employed parameters to parameters of other sounds that have similar place of articulation.* Since the majority of speech sound lies in the flat tongue space, the assignments are usually referred to the flat-tongue space [Figure 6.3].

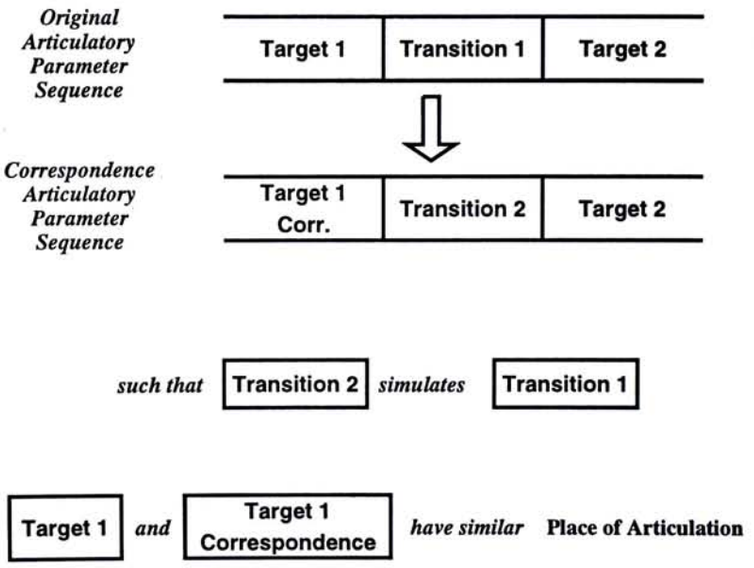

Figure 6.3: The basic idea of voice correspondence. The correspondences for segment targets enable good perceptual transition for synthesized speech signals.

Voice correspondence gives an efficient and effective way to produce transitions between segments. In human perception of speech, the quality is not only determined by stationary properties, but also depends on the properties of transitions. By using voice correspondence, the acoustic properties in the transition regions can be reliably modeled, which in turns is determined by the place of articulation of the phone units. It provides pertinent information for the interpolation of articulatory control parameters to produce good quality transitions of acoustic properties across segments. In this particular implementation, voice correspondences are mainly assigned to consonants.

6.4.1 For Nasal Sounds – Inter-Network Correspondence

To facilitate the use of parallel synthesizer networks [Section 4.2.4], voice correspondence can be used to provide guidance for the transitions of acoustic properties across different synthesizer networks. To switch between different networks, transition of the articulatory control parameters should be properly controlled to simulate the corresponding transition in acoustic properties.

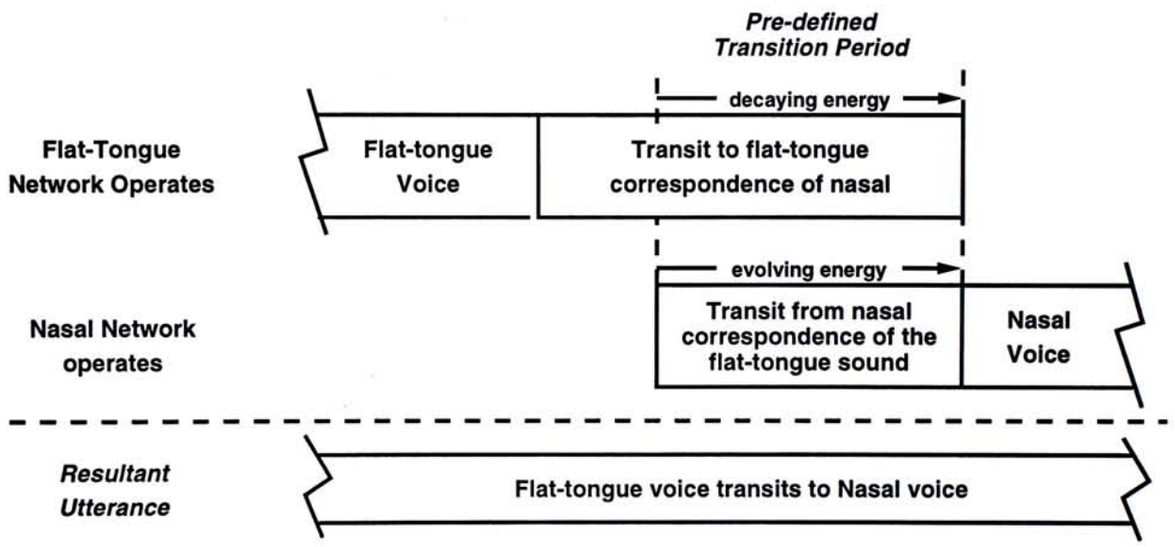

Figure 6.4: The figure shows a flat-tongue voice to nasal voice transition (e.g. for nasal codas). The flat-tongue synthesizer operates first and then decays in a short transition period. From the start of the transition period, the nasal network begins to operate. It gradually increases and finally completely takes over. The system then generates nasal sound only. For nasal to flat-tongue transition (e.g. nasal initials), the operation is similar with the operation order of synthesizer networks reversed.

In Cantonese, nasal consonants are good examples that satisfy the requirement for using parallel synthesizer networks. For example, a flat-tongue voice synthesizer network can be operated in parallel with a nasal voices synthesizer network because the transitions across flat-tongue voices and nasal voices (both initials and codas) are found to be very short. In human speech production, these transitions are usually achieved by fast flapping of articulators that add or remove the coupling to nasal tract. Hence, parallel synthesizer network is most suitable for production of speech signals in these cases.

In general, the nasal voices in Cantonese (/n/, /m/, /ng/) correspond to three different places of articulation – alveolar, labial and velar. The assignment of voice

correspondence to nasal consonants is based on these positions and the correspondence assignment is summarized in table 6.1.

Nasal Space	**Flat-Tongue Space**	**Remarks on Place of Articulation**
/n/ (1.6,1.2)	*/i/* (1.6,1.2,0.0)	Alveolar
/ng/ (0.0,0.0)	*/c/* (0.0,1.0,1.0)	Velar
/m/ (0.0,1.0)	*/u/* (0.0,1.0,1.0)	Labial is approximated by articulation position of rounded lip close vowel

Table 6.1: The correspondence assignment for the Cantonese nasal initials (and codas) to targets in flat-tongue space.

Since there must be a finite duration to change the place of articulation, the articulators position of the vowel should change gradually to approach that of the nasal consonant. The assigned voice correspondences can assist to determine the articulatory control parameters trajectory that simulates these articulator position changes for the transition.

As shown in figure 6.4, the control parameters of the flat-tongue network change slowly to the correspondence for the nasal. Meanwhile, the articulatory parameters for the nasal network emerge from the correspondence for the vowel to the targets for the nasal sound. Within the transition, there is a predefined short transition period where the signals from the two networks are overlapped and added to produce the desired vowel-nasal transition. Moreover, the energy of the flat-tongue network decreases while that of the nasal network increases gradually.

In this type of transitions, the role of the voice correspondence is to facilitate the production of a perceptually natural transition based on the concept of place of articulation. Since the parameters do not change abruptly, the effects of voice correspondence extend beyond the predefined transition period. They play an important role in producing good quality output from the flat-tongue network.

Similarly, the same approach can be applied to nasal-vowel transitions in syllables with nasal initial consonants. In figure 6.5–6.6, the control parameters profiles are shown for syllables with nasal initials and/or codas. These show the example control parameters trajectories for the nasal-vowel transitions.

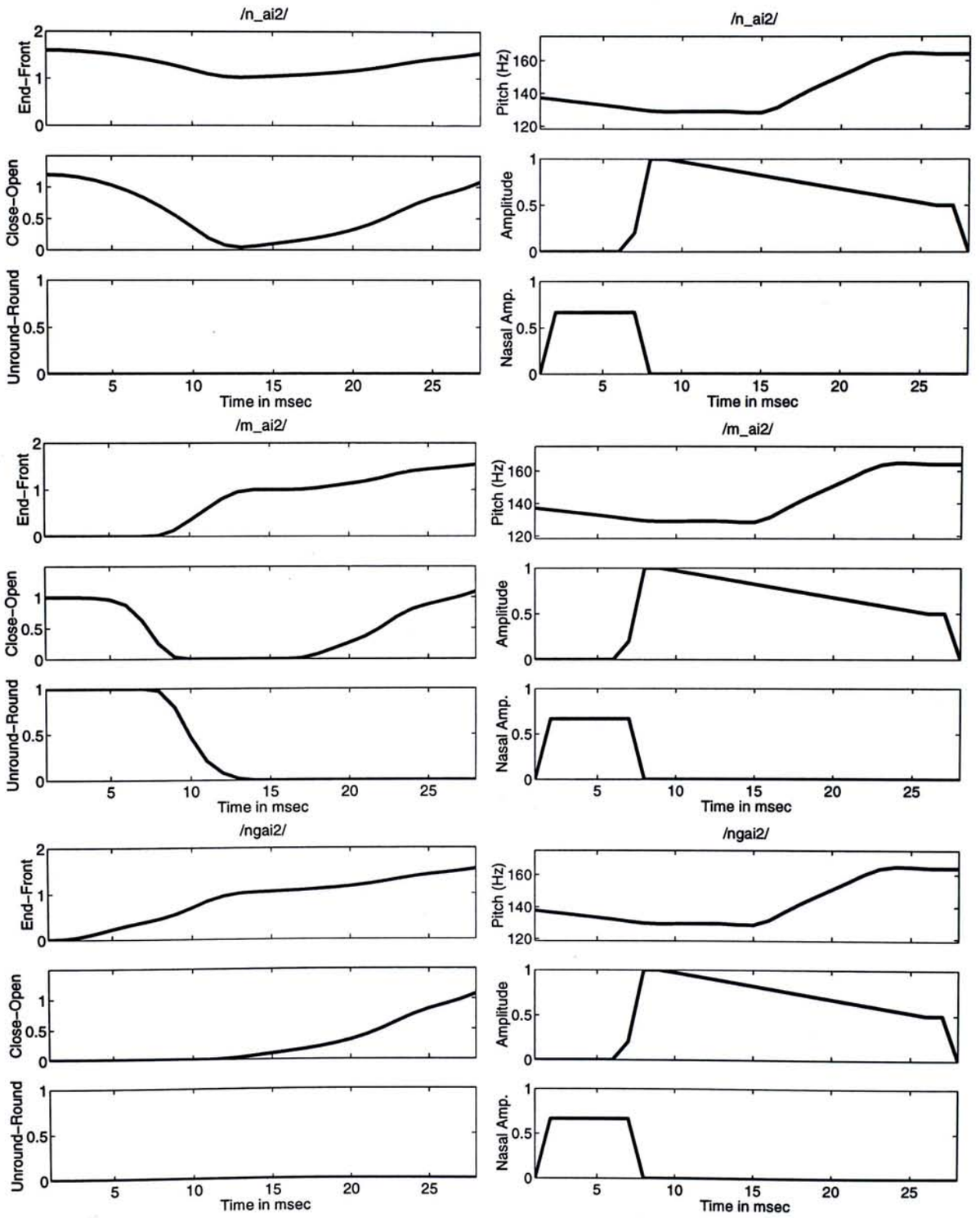

Figure 6.5: Control parameters profile for syllables (/n_ai2/, /m_ai2/ and /ngai2/) with nasal initial.

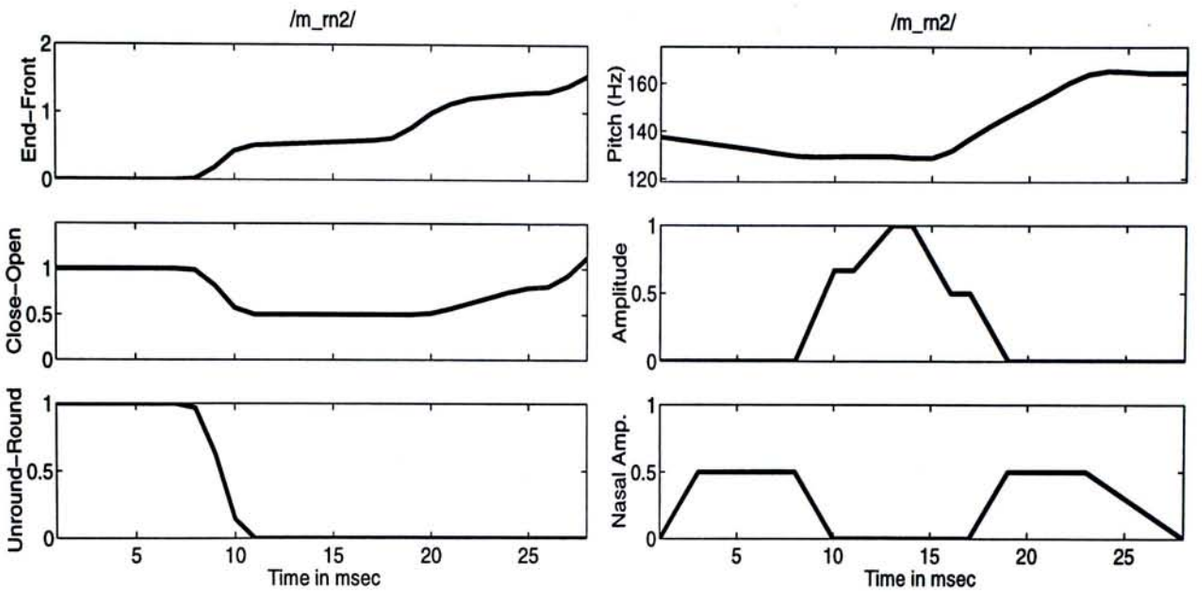

Figure 6.6: Control parameters profile for syllable (/m_rn2/) with nasal initial and coda.

In these figures, the left-hand plots are that of the articulatory control parameters amd the right-hand ones are the prosodic control of amplitude and pitch levels. These control parameters are the output of the parameter targets interpolation shown in figure 5.1. These parameters are the input fed to the synthesizer network and the IFFT operation for the final speech output.

6.4.2 In Flat-Tongue Space – Intra-Network Correspondence

In order to facilitate synthesis of different initial and coda consonants, voice correspondence is also applied within a single synthesizer network. It can significantly simplify the synthesizer network architecture. Voice correspondences for initials and codas in flat-tongue space are described in the following paragraphs.

Stop Codas

In Cantonese, there are only two types of codas : nasal and stop consonants. Nasal codas have been already discussed and only stop codas will be detailed in this part. The stop consonant codas in Cantonese are always associated with the entering tones. Moreover, there are only three types of these codas with different places of articulation : labial **/p/**,

alveolar /t/ and velar /k/. All of them are unaspirated unvoiced consonants. Their correspondences in the flat tongue space is listed in table 6.2. Figure 6.7 illustrates the basic idea of applying voice correspondence to stop codas.

Stop Coda	Flat-tongue Correspondence		Remarks on Place of Articulation
/t/	*/i/*	(1.6, 1.2, 0.0)	Alveolar
/k/	*/c/*	(0.0, 0.0, 1.0)	Velar
/p/	*/u/*	(0.0, 1.0, 1.0)	Labial is approximated by articulation position of rounded lip close vowel

Table 6.2: Flat tongue space correspondence for the stop codas in Cantonese

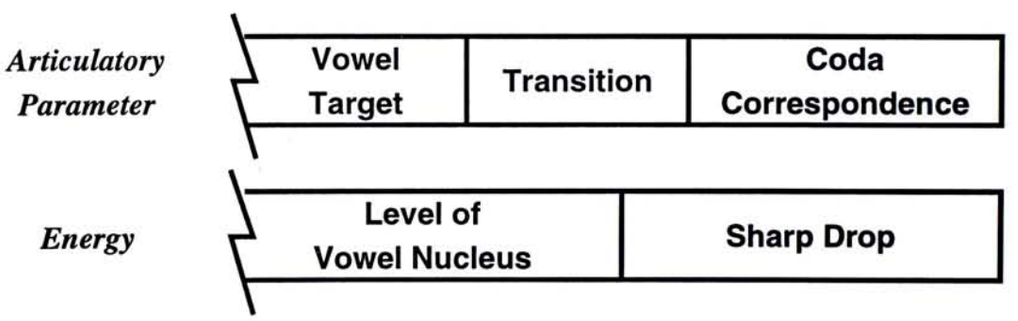

Figure 6.7: The realization of stop codas through voice correspondence. The correspondences for segment targets enable good perceptual quality for synthesized stop codas.

These stop codas always go with the entering tone and have characteristic short segmental duration. Syllables with stop consonants are realized by short duration and a sharp drop in energy for the stop effect. The assignment of correspondence is an additional feature that characterizes the stop codas by providing transition targets for the control parameters of preceding vowels. This improves the perceptual quality by proper transition of the place of articulation which in turn produce the required transition in acoustic properties. Figure 6.8 shows typical parameters trajectories for the syllables with the three stop codas and illustrates the transitions as voice correspondences.

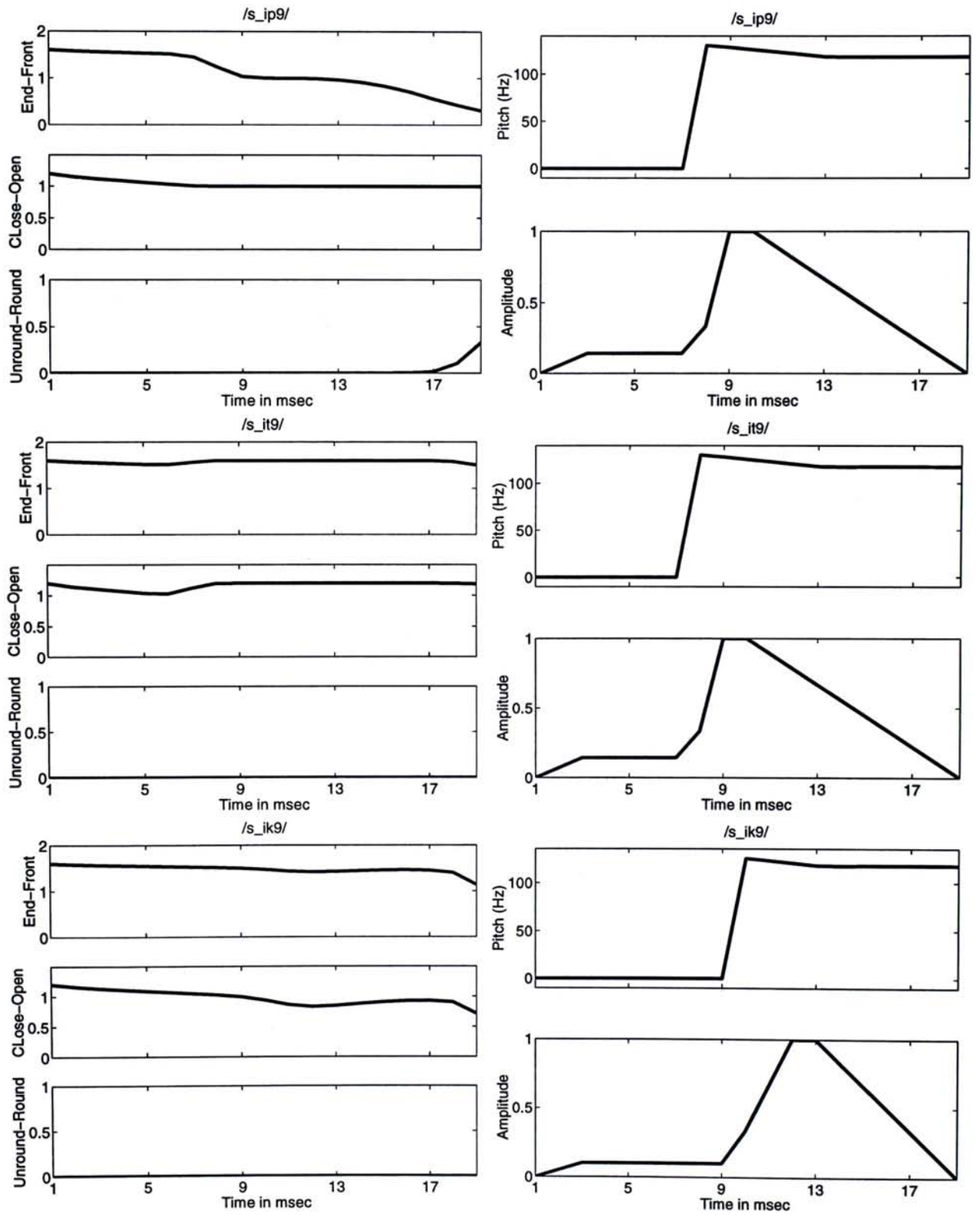

Figure 6.8: Control parameters profile for syllables (/s_ip9/, /s_it9/ and /s_ik9/) with stop coda /p/, /t/ and /k/ respectively.

Initials

In Cantonese, the number of initial consonants is larger than that of codas. The realization of these initials is determined by several factors including segmental duration and energy profile that are controlled by rules, and articulation and spectral properties that are represented by the articulatory control parameters. Since the place of articulation of most initials are similar to many corresponding flat-tongue vowels, voice correspondence is an efficient technique for producing these initials. Figure 6.9 – 6.11 show the control parameters profile of some of the Cantonese initial consonants, and table 6.3 lists out the correspondence of all Cantonese initials.

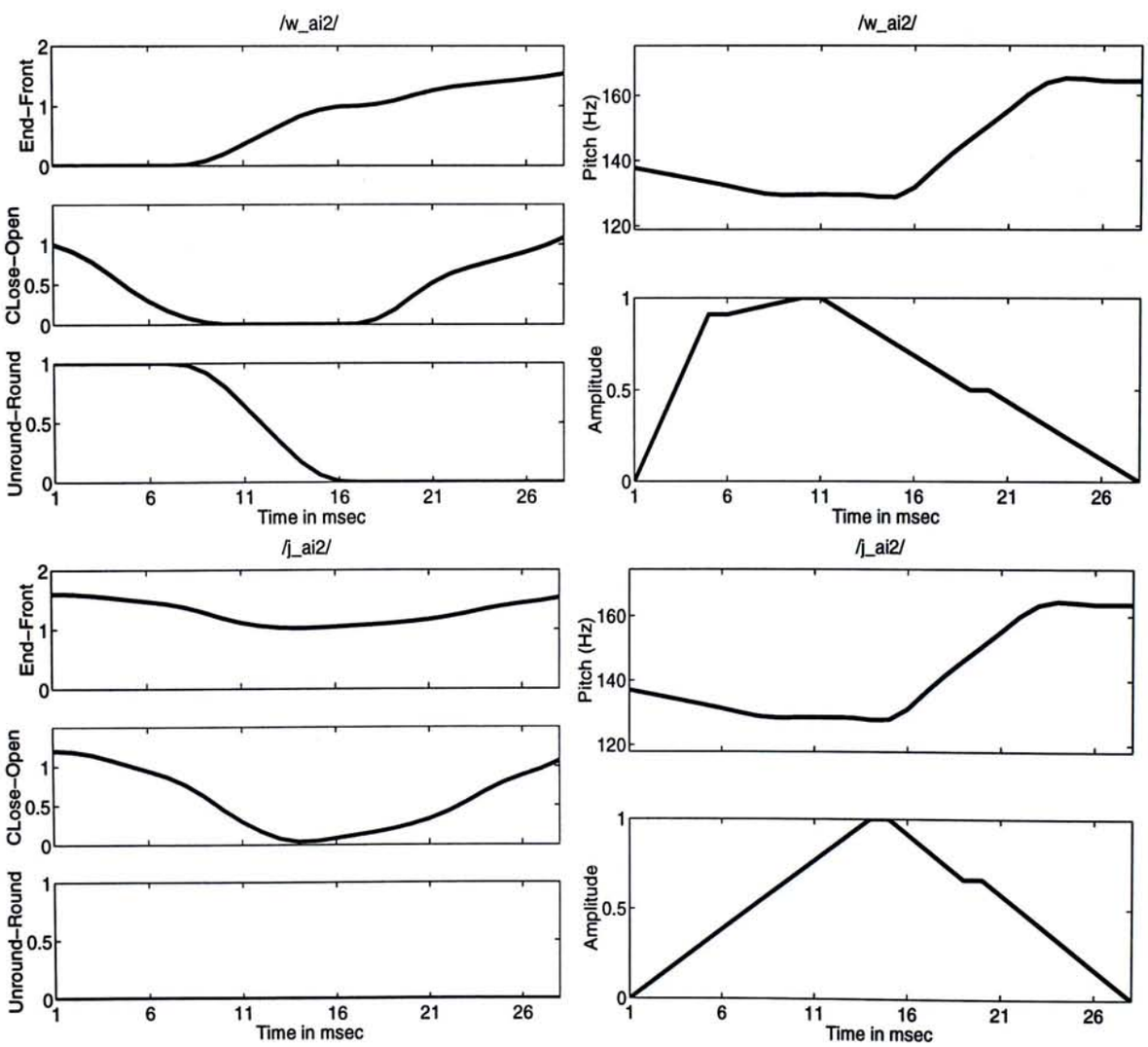

Figure 6.9: Control parameters profile for syllables (/w_ai2/ and /j_ai2/).

In figure 6.10, we can see that the place of articulation for the initial /d/ and /t/ are the same. The difference for them lies in the duration of the low energy level noisy segment between the pulsive start of the consonant and the vowel segment in the initial /t/. This low energy segment represents the aspiration section of the aspirated initial /t/.

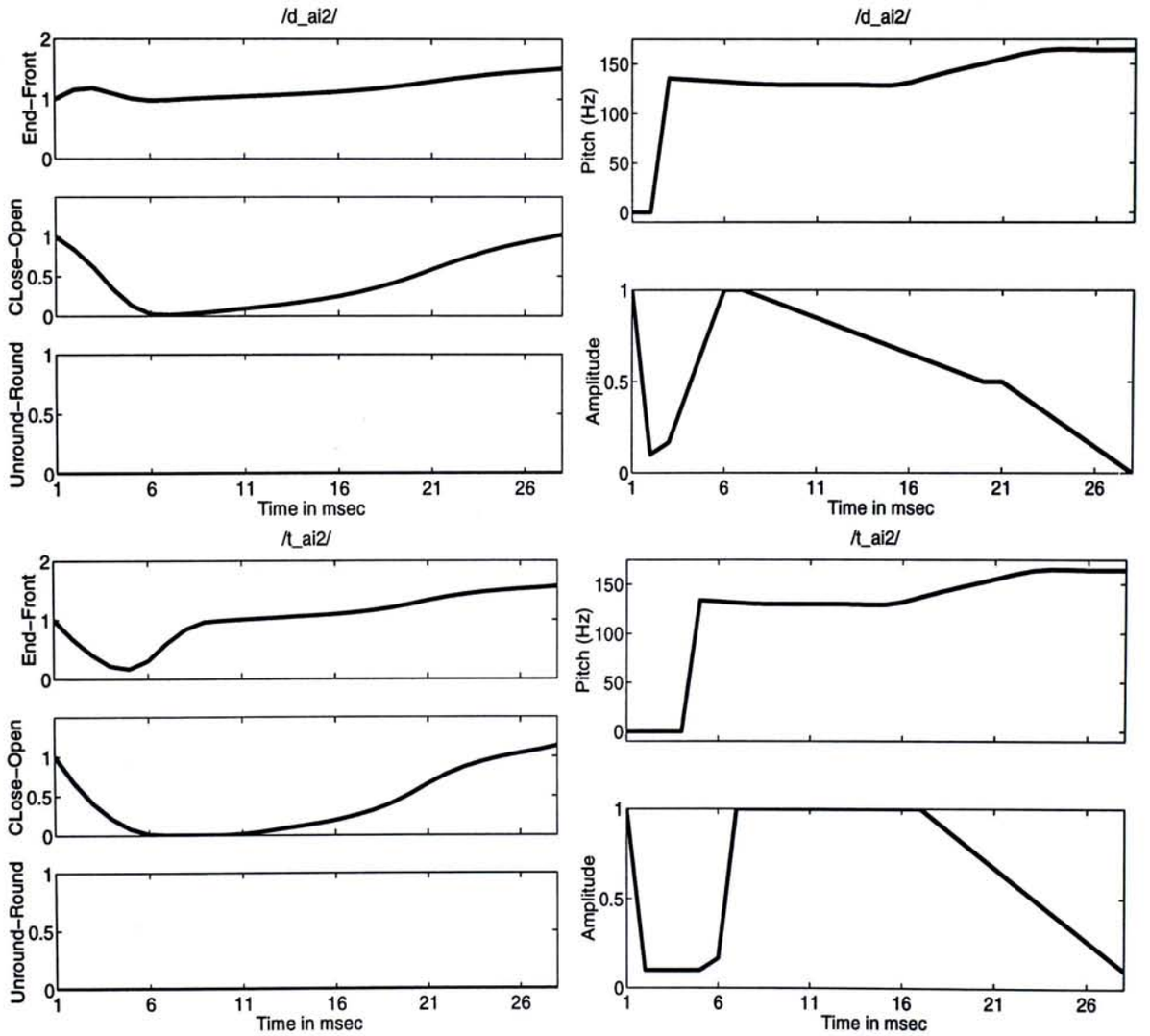

Figure 6.10: Control parameters profile for syllables (/d_ai2/ and /t_ai2/).

For the initial consonants /dz/ and /ts/, they are distinguished from the initials /d/ and /t/ by a longer pulse of energy, that is, these initials have a longer duration. On the other hand, from figure 6.11, it can be seen that the initials /dz/ and /ts/ are identified by the relatively longer and lower energy noisy region at the beginning of

/ts/ which represents the aspiration. In general, the assignment of the energy profile for realization of the initials are based on observation from the real recorded syllables. Details of realization for other Cantonese initials are tabulated in table 6.3.

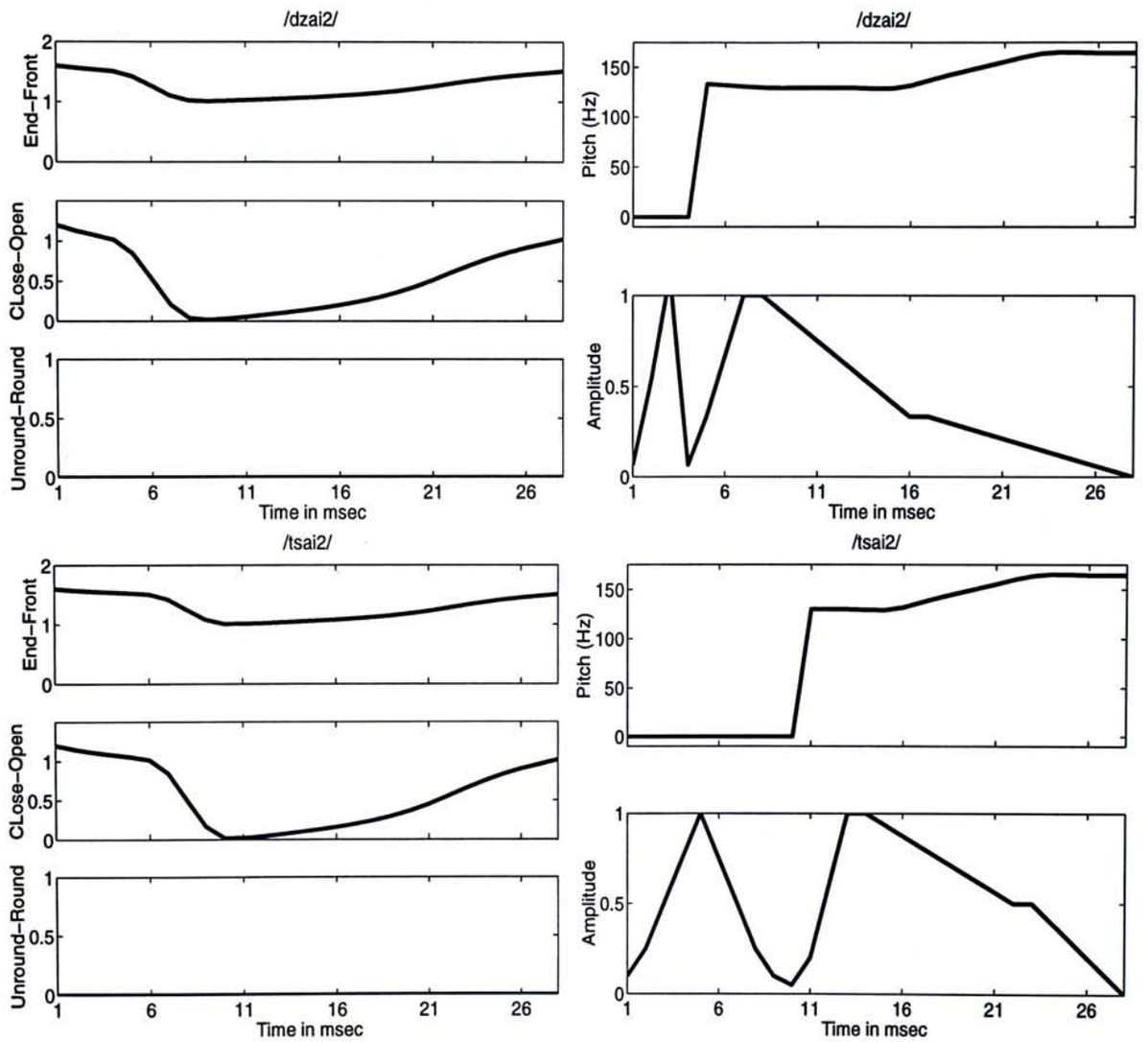

Figure 6.11: Control parameters profile for syllables (/dzai2/ and /tsai2/).

Initial	Correspondence	Remarks
/n/	*/i/*	With parallel operating nasal synthesizer network.
/m/	*/u/*	With parallel operating nasal synthesizer network.
/ng/	*/c/*	With parallel operating nasal synthesizer network.
/t/	*/i/*	Noise excited. With pulsive starting energy and a segment of low energy noise transition to vowels.
/d/	*/i/*	Noise excited. With pulsive starting energy and quick transition to vowels.
/ts/	*/i/*	Noise excited. With longer pulsive starting energy and a segment of low energy noise transition to vowels. Correspondence may change to lip-rounded */y/* for lip-rounded vowel nucleus.
/dz/	*/i/*	Noise excited. With longer pulsive starting energy and quick transition to vowels. Correspondence may change to lip-rounded */y/* for lip-rounded vowel nucleus.
/p/	*/u/*	Noise excited. With pulsive starting energy and a segment of low energy noise transition to vowels.
/b/	*/u/*	Noise excited. With pulsive starting energy and quick transition to vowels.
/k/	*/c/*	Noise excited. With pulsive starting energy and a segment of low energy noise transition to vowels. The correspondence changes with the vowels according to rules.
/g/	*/c/*	Noise excited. With pulsive starting energy and quick transition to vowels. The correspondence changes with the vowels according to rules.
/w/	*/u/*	With gradual increasing energy in association with the transition of articulatory parameters to vowel target.
/j/	*/i/*	With gradual increasing energy in association with the transition of articulatory parameters to vowel target.
/kw/	*/c/ - /u/*	Combination of */k/* and */w/* with additional transition inside the initial and modified timing.
/gw/	*/c/ - /u/*	Combination of */g/* and */w/* with additional transition inside the initial and modified timing.
/s/	*/i/*	Noise excited. With a low energy level segment. Correspondence may change to lip-rounded */y/* for lip-rounded vowel nucleus.
/h/	dynamic	Noise excited. The articulatory parameters depend on the vowel nucleus of the syllable and the preceding voice segment, if exists. The articulatory control vector is a transition from the preceding to the targeted vowel.
/f/	*/u/*	Concatenated from stored templates. Make use of the correspondence for transition.
/l/	dynamic	Concatenated from stored templates.

Table 6.3: All Cantonese initials are included in the table showing their correspondence in flat-tongue space together with the implementation remarks.

6.5 Summary

In summary, the proposed articulatory control provides an alternative approach for the synthesis control in speech synthesis. It provides better flexibility and controllability. At the same time, this intuitive approach for synthesis also allows much logical simplification for synthesis control based on the knowledge of actual human speech production mechanism. These simplifications includes parallel operating synthesizer network that simplifies the network architecture and the voice correspondence that applied to most consonant generations. The voice correspondence is derived from knowledge of the articulation positions in the real speech production system. It gives an efficient implementation of the synthesizer network while still provides a good synthesis control model for speech production. It also provides guidelines for spectral transition across different parallel operating synthesizer networks. All in all, articulatory control for the phone-based neural network concatenation synthesis creates much room for innovative design and enhancement of the synthesis control over traditional concatenation.

Chapter 7

Pause Duration Properties in Cantonese Phrases

To study the prosodic variations due to coarticulation in spoken Cantonese and also build up basis for the prosodic control in speech synthesis, experiments have been performed to investigate some of the underlying phenomena. In this chapter, concentration will be made on the inter-syllable pause duration in Cantonese phrases. Statistical measurement on the pause durations under different acoustic context shows that the ratio of pause duration to phrase duration is consistent among different trials on phrases of two and three/four syllables long. The result provides useful basis to determine the duration of pauses to be inserted between syllables in speech synthesis. Figure 7.1 illustrates an example phrase including phrases with inter-syllable pause and also without it. To certain extent, the existence of inter-syllable pause is believed to depend on the neighbouring phones.

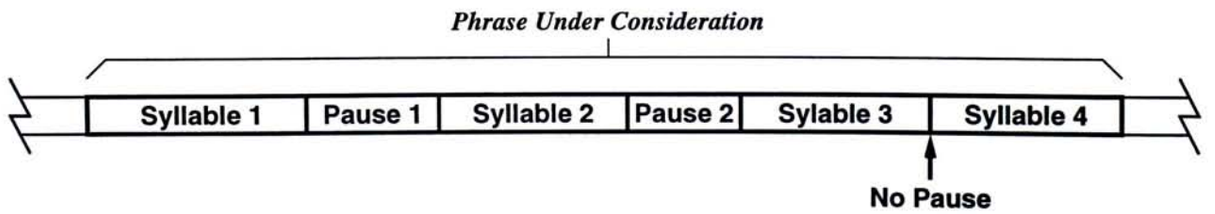

Figure 7.1: The inter-syllable pauses between the syllables in a phrase is shown. The pause durations depend on the adjacent phones. In some cases, pause may be absent and the syllables are uttered continuously without pause.

7.1 The Prosodic Feature - Inter-Syllable Pause

The inter-syllable pause duration is one of the many different prosodic features in spoken language. It is particularly important in syllabic languages, specifically for many oriental languages and dialects like Mandarin and Cantonese, where sentence meaning are based on the syllable meaning. Incorrect insertion or assignment of these features may result in a change of the perceived meaning of utterance.

Basically, the pauses between syllables have several major functions. They are added to mark acoustical boundaries for the syllables. The pause duration between pairs of syllables in phrases and that in sentences are fairly different. Perceptually, the pause durations also determine the grouping of syllables as phrases in continuous speech. In general, it assists the delivery of linguistic information in a clearer way. If the inter-syllable pause durations are set inappropriately, there exists chances that the delivered meaning or speaking attitude is altered. Or more importantly, a completely different idea might be delivered due to a misplacement of the phrase and sentence boundaries in the spoken utterance.

Furthermore, pause durations together with segmental durations control the rhythm of the utterance. Well-controlled rhythm will give better naturalness in the sense of prosody. If the pause durations and segmental durations are not properly inserted, the uttered sound will sound more mechanical even the information is intelligibly received. Prolonged listening of this type of speech may cause fatigue and stress to the users.

Certainly, the inter-syllable pause duration itself does not completely control the delivery of information in spoken language. It is the joint effects of different linguistic, prosodic and acoustic features that determine the information delivery. Each property only makes a specific contribution in a particular area.

7.2 Experiment for Measuring Inter-Syllable Pause of Cantonese Phrases

Several tests have been conducted to examine the pause duration distribution of Cantonese. These tests mainly focus on the inter-syllable pause duration between syllables in two-syllable to three/four syllable phrases. This particular timing characteristic governs the fluency and to a certain degree, the naturalness of continuously uttered phrases.

In this experiment, the basic assumption is that the effect of one segment does not spread over to more than one of its adjacent segments. It determines the pause duration right next to it but not extend any longer. The validity of this assumption is consolidated by the required fluency in the uttered samples which makes the physical limitation of articulator positions dominant in the determination of the pause durations.

7.2.1 Speech Material Selection

The data set used in the tests are chosen so as to cover most of the possible initial to final phone-pairs. In the selection process, the lexical tone of the syllable is ignored. It is expected to have minor effect on the pause duration since a large portion of the involved phones are unvoiced consonants. In addition, null initial is not included and rare combination of phone pairs are also excluded. To be more specific, the */j_/* vowel final (an extremely rare vowel final in Cantonese) is excluded in the list of phrases since there is not many commonly used phrases.

On the other hand, diphthongs with the same final phones are put to the same class, e.g. */ei/*, */ai/*, */ci/* and */au/*, */ru/*, */cu/* are in the same final phone class. For the initial consonants, similar grouping is also made. */g_/*, */gw/* and */k_/*, */kw/* are put into the same initial phone class respectively.

The list of three/four-syllable phrases is constructed by extending the two-syllable phrases in the list to phrases of four syllables. Basic criteria of phrase selection is listed in table 7.1. The complete lists of the two-syllable phrases and three/four-syllable phrases are given in Appendix D.1 and Appendix D.2 for reference purpose.

It should be noted that since the three/four-syllable phrases are made up from the set

of two-syllable phrases, therefore, there is always a corresponding pause in the three/four-syllable phrase for the two-syllable phrases. This enables the investigation of the effect of phrase duration on the inter-syllable pause duration. It also allows easier and more reasonable comparison to be made among the pause-ratio's from phrases of different syllable count. Hereafter, pauses in three/four syllable phrases refer specifically to these corresponding pauses.

Exclusion	Lexical tone of syllable is ignored
	Null initial is not counted as initial
	Final vowel */j_/* is excluded
Inclusion	All possible pairs of initial and final phone-pairs
	Diphthongs with same final phones are put to same phone class
	Initial consonants with same starting phone are put into same phone classes

Table 7.1: Summary of the criteria in the selection of phrases for the two-syllable and three/four-syllable phrase list.

7.2.2 Experimental Procedure

Speakers are invited to read out the lists of phrases – 238 two-syllable phrases and 238 three/four-syllable phrases. The speakers are required to utter the list of phrases naturally and it is not necessary to exaggerate the clearness of individual syllables.

The speech data is recorded using Technics RS-B50 tape recorder. A/D conversion is made using the aux line of Sun Sparc DBRI audio interface. The sampling rate is 16k Hz at 16 bit signed integer precision.

There are three speakers invited to the experiment – 1 male and 2 female. It makes up a total of eight samples. The speaker information is summarized in table 7.2.

Speaker ID	m01	f01	f02
Gender	male	female	female
Trial	one	two	five

Table 7.2: Speaker information of recorded phrase lists for pause duration measurement.

In each trial, there are 913 pauses included [1]. Therefore, the recorded speech provides a total of 7304 pauses from the 8 trials giving rise a moderate sample size for the experiment.

Measurement and Analysis

The tape-recorded speech is digitized and manually labeled to mark the boundaries of syllables and phrases. Based on the marked boundaries, the inter-syllable pause duration and the total phrase duration are computed. Comparison is made on the ratio of the pause duration to the phrase duration.

Several measurements have been derived and analyzed from the computed result.

1. Mean and standard deviation of the pause duration for verifying the consistency of measured result in different trials

2. Ratio of pause duration to phrase duration is computed as pause-ratio

$$pause\ ratio = \frac{pause\ duration}{phrase\ duration}$$

3. Means of pause-ratio in two-syllable phrases and means of pause-ratio in three/four-syllable phrases are compared

4. Mean of normalized pause-ratio are calculated by multiplying the syllable count in the phrases to mean pause-ratio

$$normalized\ pause\ ratio = pause\ ratio \cdot syllable\ count\ in\ phrase$$

5. The ratio of pause-ratio in 2-syllable phrase to 3/4-syllable phrase is defined as a pause-ratio factor for investigation of the change in pause due to phrase length.

$$pause\ ratio\ factor = \frac{two\ syllable\ phrase\ pause\ ratio}{three/four\ syllable\ phrase\ pause\ ratio}$$

[1] 238 from 238 two-syllable phrases, 39x2 from 39 three-syllable phrases and 199x3 from 199 four-syllable phrases.

7.2.3 Result

Table 7.3 and table 7.4 shows the mean pause-ratio of two-syllable Cantonese phrases and the mean pause-ratios of the corresponding pauses in three/four syllable Cantonese phrases respectively. Preliminary result shows general consistency in the pause-ratio, and detail analysis will be given in the following sections.

Figure 7.2 shows the pause-ratio factor for two-syllable phrases to corresponding pause in three/four syllable phrases. The syllable count normalized ratio distribution is also given.

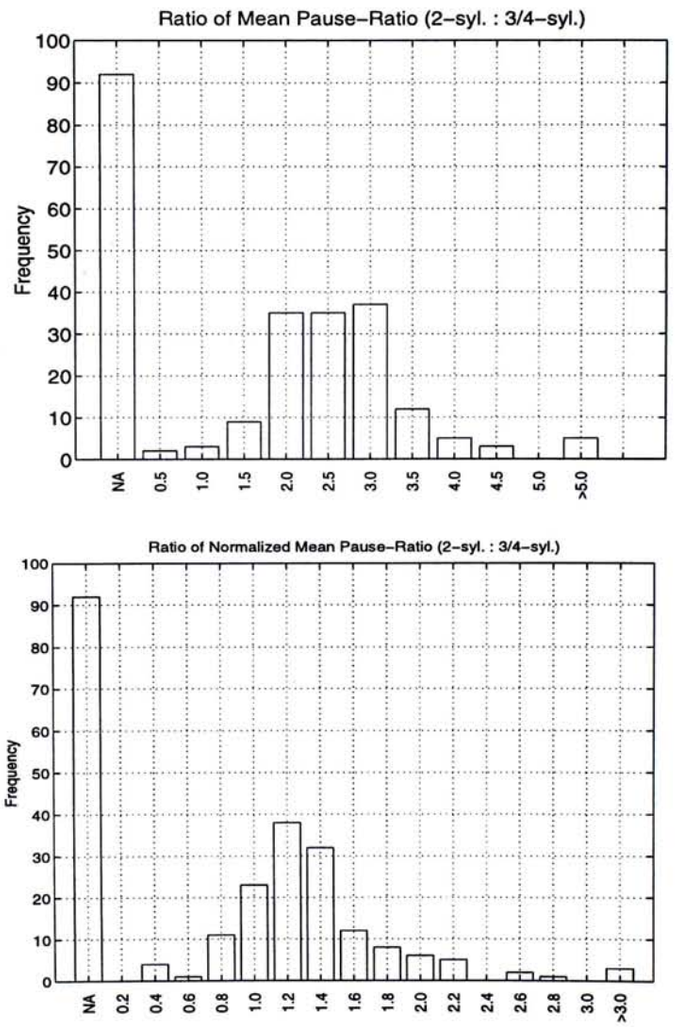

Figure 7.2: The distribution of the pause-ratio factor of two-syllable phrases to the corresponding pause in three/four syllable phrases. Clustering of the factor is observed and the clustering increases in the normalized case. It shows that there is a slight but consistent decrease of pause duration in longer phrases.

In pairs of /X last/-/next X/

last \ next	/j-/	/s-/	/t-/	/d-/	/ts/	/n-/	/dz/
/i/	0	0	0.08396	0.10294	0.02824	0	0.05822
/e/	0	0	0.05679	0.06817	0.03997	0	0.07529
/y/	0	0	0.06635	0.06724	0.04557	0	0.07951
/j_i_/	0	0	0.05263	0.08955	0.02577	0	0.07015
/u/	0	0	0.06292	0.10351	0.03404	0	0.04126
/a/	0	0	0.06548	0.12758	0.05546	0	0.05301
/c/	0	0	0.05702	0.08429	0.03360	0	0.05485
/w_u_/	0	0	0.04870	0.09210	0.03729	0	0.08030
/_n/	0	0.00120	0.01233	0.04553	0.00240	0	0.04531
/_5/	0	0	0.02882	0.04696	0.00417	0	0.03185
/_m/	0	0.00168	0.03653	0.07111	0.03113	0	0.02260
/_p/	0.22887	0.13196	0.28207	0.41106	0.29822	0.20399	0.32999
/_t/	0.26117	0.02666	0.22064	0.28497	0.15919	0.24102	0.18781
/_k/	0.21193	0.12020	0.16010	0.32223	0.35765	0.22152	0.28875

last \ next	/p-/	/w-/	/m-/	/b-/	/f-/	/h-/	/l-/	/g-/	/k-/	/ng/
/i/	0.06752	0	0	0.12581	0	0	0	0.09092	0.06278	0.00567
/e/	0.06716	0	0	0.09052	0	0	0	0.08962	0.06115	0.00225
/y/	0.05693	0	0.00681	0.09598	0	0	0	0.08959	0.07191	0.00257
/j_i_/	0.06044	0	0	0.10820	0.01099	0	0	0.08935	0.03776	0.00548
/u/	0.05341	0	0	0.09327	0	0	0	0.14718	0.05461	0.00601
/a/	0.07435	0	0	0.10661	0	0	0	0.11722	0.08952	0.00485
/c/	0.09875	0	0	0.07956	0	0	0	0.10109	0.12573	0
/w_u_/	0.07824	0	0	0.09713	0	0	0	0.07260	0.08619	0.00769
/_n/	0.03544	0	0	0.04977	0.00669	0	0	0.03553	0.02708	0.01432
/_5/	0.01514	0	0	0.05510	0.04230	0	0	0.03862	0.02778	0.00853
/_m/	0.03580	0	0	0.05691	0	0.00537	0	0.03841	0.04397	0.01805
/_p/	0.22542	0.19014	0.20990	0.46440	0.10977	0.19783	0.22833	0.36440	0.26656	0.14231
/_t/	0.18784	0.30079	0.17139	0.30272	0.13383	0.13464	0.14181	0.25902	0.21915	0.18729
/_k/	0.23465	0.30793	0.17248	0.23704	0.07599	0.15219	0.34709	0.21313	0.27830	0.14575

Table 7.3: Mean values of pause-ratio in two-syllable Cantonese phrases.

In pairs of /X last/-/next X/

last \ next	/j-/	/s-/	/t-/	/d-/	/ts/	/n-/	/dz/
/i/	0	0	0.04040	0.05055	0.01045	0	0.03207
/e/	0	0	0.05276	0.06626	0.02945	0	0.04285
/y/	0	0	0.02953	0.03778	0.01392	0	0.04942
/j i-/	0	0	0.01964	0.04569	0.01972	0	0.03596
/u/	0	0	0.02184	0.04483	0.01952	0	0.02597
/a/	0	0	0.02083	0.04818	0.02231	0	0.02932
/c/	0	0	0.01335	0.04003	0.01963	0	0.02289
/w u-/	0	0	0.01649	0.06068	0.01915	0	0.03251
/_n/	0	0	0.00152	0.01564	0.00500	0	0.01302
/_5/	0	0	0.01203	0.01607	0.00844	0	0.01349
/_m/	0	0	0.01391	0.02829	0.00830	0	0.00570
/_p/	0.05536	0.07457	0.17622	0.12680	0.10553	0.06238	0.12114
/_t/	0.07050	0	0.08040	0.10665	0.06618	0.04778	0.06316
/_k/	0.08041	0.04707	0.08558	0.15244	0.14028	0.09011	0.10784

last \ next	/p-/	/w-/	/m-/	/b-/	/f-/	/h-/	/l-/	/g-/	/k-/	/ng/
/i/	0.01300	0	0.00170	0.05748	0	0	0	0.04535	0.03496	0
/e/	0.02653	0.00393	0	0.04658	0	0	0	0.05289	0.05165	0
/y/	0.03592	0	0	0.04153	0	0	0	0.04240	0.03261	0.00462
/j i-/	0.03541	0.00436	0	0.06882	0.01509	0	0	0.05672	0.02657	0.00586
/u/	0.02307	0	0	0.05718	0	0	0	0.05695	0.03035	0
/a/	0.02963	0	0	0.06872	0	0	0	0.05673	0.05805	0
/c/	0.04612	0	0	0.04487	0	0	0	0.04794	0.05835	0.00543
/w u-/	0.03221	0	0	0.05759	0	0.00862	0	0.03320	0.02960	0.00581
/_n/	0.01313	0	0	0.01933	0.00180	0	0	0.03396	0.01380	0.01070
/_5/	0.00696	0	0	0.02575	0.01672	0	0	0.01763	0.00929	0
/_m/	0.01403	0	0	0.02438	0	0	0	0.01680	0.02485	0.00585
/_p/	0.10358	0.12391	0.07229	0.16995	0.03411	0.07667	0.07591	0.16952	0.10123	0.05829
/_t/	0.09301	0.08675	0.06723	0.11388	0.02402	0.01973	0.05549	0.13366	0.08098	0.05631
/_k/	0.14554	0.13423	0.05376	0.12067	0.01890	0.05090	0.09042	0.12695	0.10525	0.04750

Table 7.4: Mean values of pause-ratio in three/four-syllable Cantonese phrases.

7.3 Characteristics of Inter-Syllable Pause in Cantonese Phrases

7.3.1 Pause Duration Characteristics for Initials after Pause

Based on the initial after pause (if exists), the following characteristics are observed

- There is no inter-syllable pause before the following initials : ***/j_/***, ***/w_/***, ***/l_/***, ***/n_/***, ***/m_/***, ***/f_/***, ***/h_/*** and ***/s_/***. It can be seen that these initials belong to either of these initials categories : glides, liquids, nasals, and fricatives.

- There are exceptions particularly when the preceding syllable is in entering tone. In these cases, there are inter-syllable pauses due to the preceding final codas dominant.

- The nasal initial ***/ng/*** is usually associated with a very short pause duration.

- Those initial consonants with sudden burst of energy (e.g. plosive) will have a finite period of pause.

- Initials with aspirating properties are generally associated with shorter pause than the corresponding non-aspirating counter parts. The aspirating and unaspirating pairs of initials are given in table 7.5.

  The pause-ratio against different preceding finals in the initial pairs /t_/ – /d_/,

Aspirating	**Unaspirating**
/t_/	***/d_/***
/ts/	***/dz/***
/p_/	***/b_/***
/k_/	***/g_/***

Table 7.5: The pairs of aspirating and unaspirating initials that are similar in place and manner of articulation.

/ts/ – /dz/, /p_/ – /b_/ and /k_/ – /g_/ are plotted in figure 7.3, 7.4, 7.5 and 7.6 respectively.

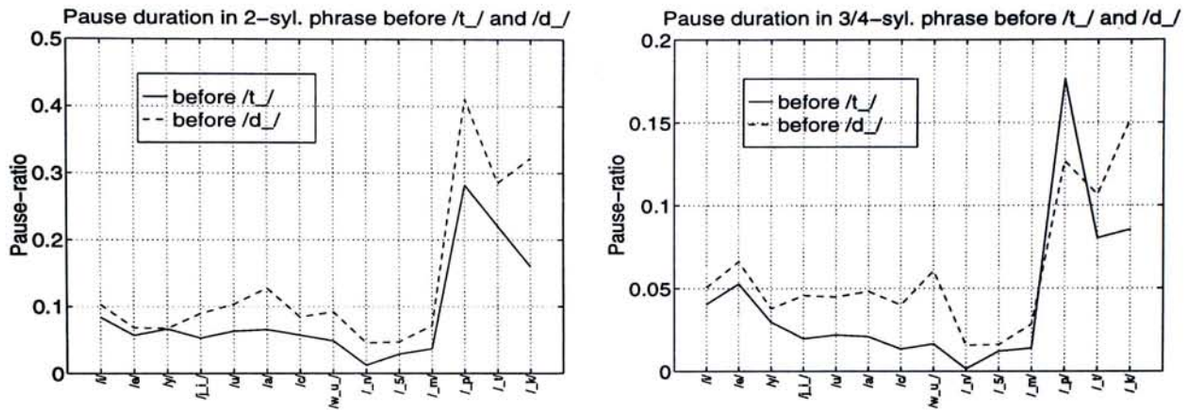

Figure 7.3: The figures compare the pause-ratio before different pause preceding finals for the initials /t_/ and /d_/.

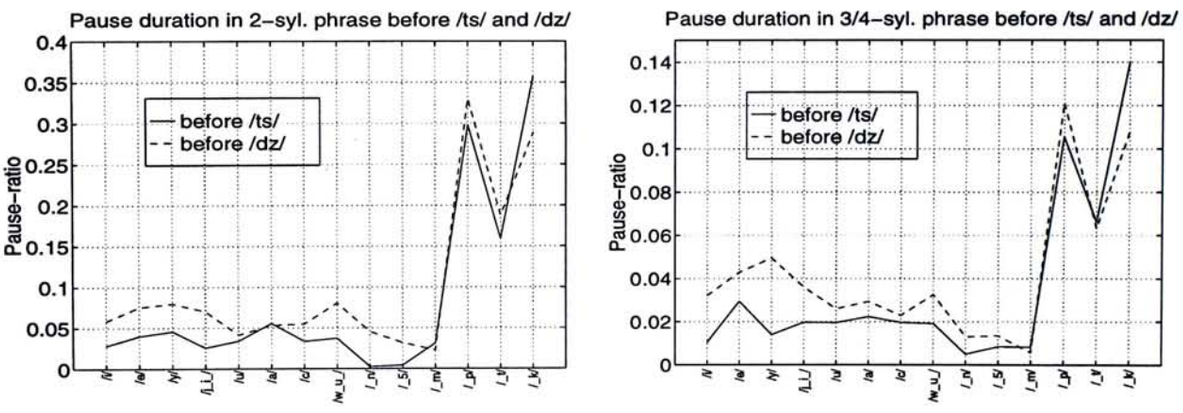

Figure 7.4: The figures compare the pause-ratio before different pause preceding finals for the initials /ts/ and /dz/.

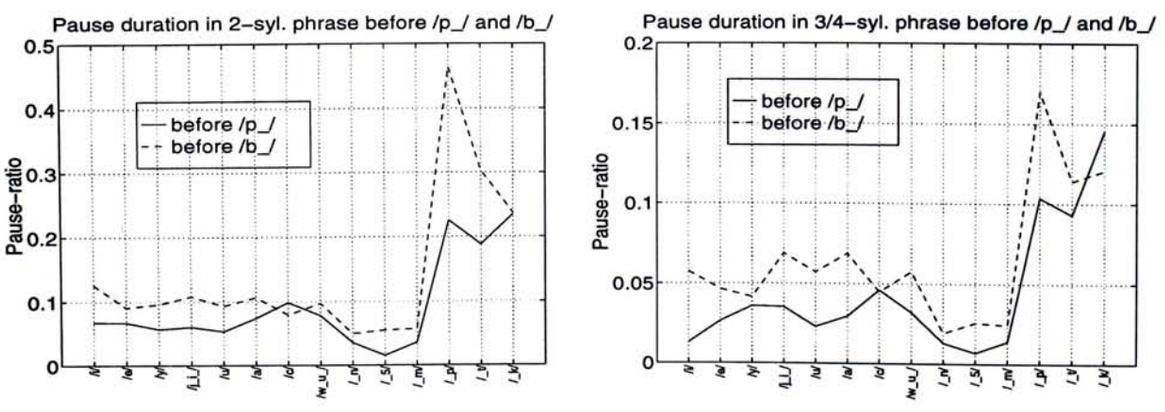

Figure 7.5: The figures compare the pause-ratio before different pause preceding finals for the initials /p_/ and /b_/.

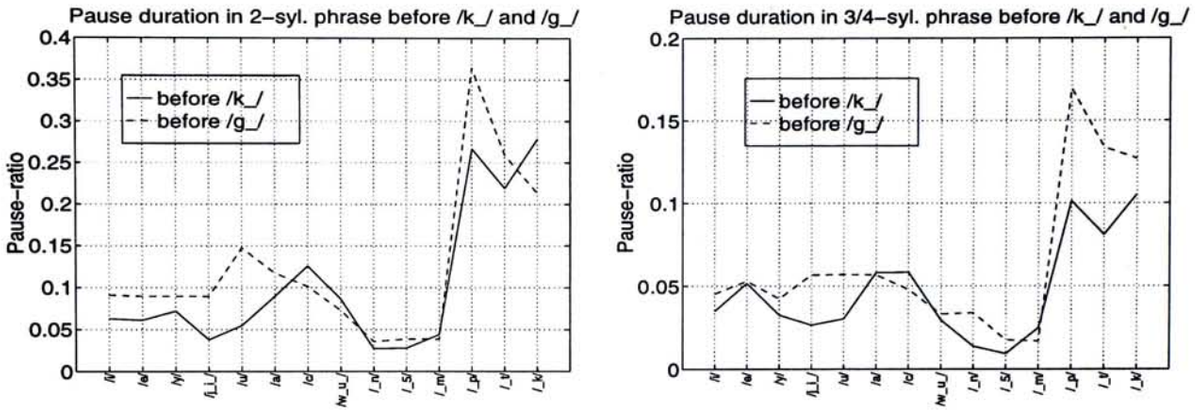

Figure 7.6: The figure compare the pause-ratio before different pause preceding finals for the initials /k_/ and /g_/.

- If the preceding syllable is in entering tone, the decrease in pause duration for aspirating initial is not necessary to be in proportion.

7.3.2 Pause Duration Characteristic for Finals before Pause

Based on the final phones of the syllables preceding pauses, the following phenomena are detected.

- In general, entering tones are associated with a pause period regardless of the initial of the following syllable.
- As mentioned earlier, syllables preceded pauses with entering tone (***/_p/, /_t/, /_k/*** finals) will have dominant pause duration. They may override the effect due to the initial of the syllable following the pause.
- In case the final coda preceding the pause is a nasal, the associated pause duration is shorter (if exists) than those with vowel (include diphthong) finals.

7.3.3 General Observations

Consistent pattern of existence of inter-syllable pause

The existence of pause is generally consistent among all speakers and trials, with very little exception. This consistent pattern is due to the fact that certain phone-pairs are not

associated with inter-syllable pause in fluently spoken phrases. Therefore, the existence of inter-syllable pauses depends only on some specific inter-syllable phone-pairs. This consistency in existence consolidates the assumption that coarticulation effect between syllables take a dominant part in the pause duration of the data recorded.

From the computed pause ratio [figure 7.3], it can be seen that pause is consistently associated with initials with pulsive energy. Those glides, liquids, nasals (except ***/ng/***) and fricatives do not have pulsive starting energy and so do not have inter-syllable pause.

General consistency in pause-ratio for all phone-pairs

The pause ratio for each phone-pair between the syllables generally has only small deviation among speakers and trials. This observed consistency ensures that the pause ratio in a phrase depends very much on the particular phone-pairs in the fluently spoken utterance.

In addition, the ***pause-ratio factors*** are found to be quite consistent across different phone-pairs and their distribution [figure 7.2] shows that there is a fairly constant normalized factor between short and long phrases. In this case, the factor is around 1.2 for two-syllable and three/four syllable phrases. This means that there is a slight decrease in the pause duration in longer phrases as expected.

Some initial consonants are usually not associated with pause

Three main categories of initials usually have no preceding pause in their inter-syllable transitions.

1) noisy consonants, ***/s_/***, ***/f_/***, ***/h_/***;

2) glides and liquids, ***/w_/***, ***/j_/***, ***/l_/***;

3) nasal consonants, ***/n_/***, ***/m_/***, except ***/ng/***

These consonants do not have pauses mainly because they do not have pulsive consonant energy. Unless the phrase is very long, the noisy consonants do not need pause period to be properly pronounced and the air-flow of the preceding final can be continued without break.

Glides and liquids do not introduce pause because they are practically transition vowels. On the other hand, acoustic coarticulatory phenomenon in the inter-syllable

transition from the preceding final phones are often observed in glides. Coarticulatory effects are usually found instead of pause in the transition to semi-vowel initials.

Based on the idea of voice correspondence, nasals are viewed as decaying voices of the correspondence points in flat tongue space. There is additional parallel exciting nasal resonance tube of increasing energy level. This means that they resemble fast transition vowels rather than sudden energy change. Therefore, there is not any inter-syllable pause in front of these initials.

An exceptional nasal is ***/ng/***, there is always a short pause before it. It is believed to be caused by the manner of articulation. This velar nasal is associated with a velar close at the start of articulation. This temporary close of the vocal tract leads to the short pauses. Moreover, the pause durations also decreases as the syllable count in the phrase increases.

Pauses in long phrases are shorter than those in short phrases

It is found that the normalized pause-ratio factor [Figure 7.2] is just above one while the raw pause-ratio factor is slightly greater than two. This indicates that the pause-ratio in 2-syllable phrases are generally longer. The normalized pause-ratio factor (1.2) shown in figure 7.2 also indicates that there is a 20% increase in the pause-ratio from 2-syllable phrases to 3/4-syllable phrases. The result also shows that there is a general consistency in this duration increment without any dramatic change.

7.3.4 Other Observations

Besides the statistics for the inter-syllable pause duration, some other characteristics have also been noticed.

Voice Correspondence of consonants in flat tongue vowel space is verified by coarticulatory transitions

It is experimentally verified that the transitions of the phone in *consonant–vowel* are perceptually similar to the phone transitions in *corresponding vowel–vowel.* This is observed through the acoustic coarticulatory transitions.

As an illustrating example, the transition for ***/XXXi/-/m_XX/*** is found to have the former syllable coda, ***/_i/*** pronounced with trailing phone, ***/u/***, becoming ***/XXXiu/-/m_XX/***. This effect is more prominent in the ***/i/*** and ***/u/*** corresponding initials such as ***/w_/, /j_/, /m_/, and /p_/*** etc. This observation well agrees with the idea of voice correspondence that is based on the place of articulation of sounds.

Acoustic coarticulatory effect may extend across pause

The acoustic coarticulatory effect may extend across the inter-syllable pauses. That is, pause may lie between the phone-pairs while the acoustic properties are carried across the pause to the affected phone.

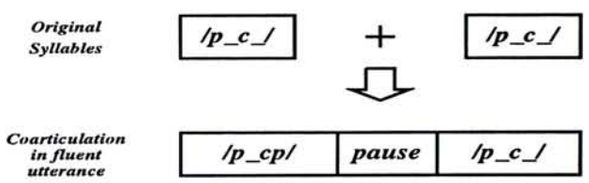

Figure 7.7: The figure is an example showing the coarticulatory effect that extends across the inter-syllable pause.

As shown in the example of figure 7.7, the phrase ***/p_c_/ /p_c_/*** is found to be uttered as ***/p_cp/ /pause/ /p_c_/***. This type of acoustic coarticulatory effect across pause is also found in plosives such as ***/p_/, /b_/, /d_/, /t_/, /dz/, /ts/***, etc.

/h_/ is highly influenced by its surrounding phones

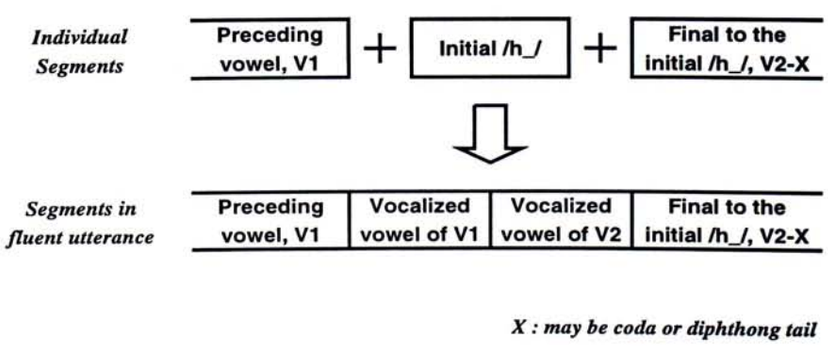

Figure 7.8: The assimilation of the initial /h_/ by its adjacent phone units which are not necessary its final.

The consonant */h_/* is found to be the most easily affected one. If it is started from total silence, it has a clear noise-like waveform. However, if it appears as the initial of non-starting syllables in phrases, it is will depend on the final phone of the previous syllable [Figure 7.8]. The more fluently the phrases are uttered, the more it is affected, even to the extent that the noise-like property becomes very minor and the vocalized noisy version of the preceding and following phones become dominant in the "consonant". Another prominent observation of the consonant */h_/* is that unless the phrase is intentionally separately uttered, there is usually no inter-syllable pause involved except after the entering tone (i.e. coda */p_/*, */t_/* and */k_/*).

7.4 Application of Pause-duration Statistics to the Synthesis System

The statistics of pause duration is found to have certain degree of consistency. These data are usually found useful in concatenation of synthesized syllables or just recorded templates. Based on the the length of the whole phrase, appropriate duration of pause is inserted as a fraction of the total phrase duration.

The basis of pause duration timing rules can be derived from the two tables of mean pause-ratio's [Figure7.3 and 7.4]. Insertion of proper pause between syllables in phrases will enhance the perceived rhythm of the utterance.

In the prototype system, incorporation of this pause insertion information to the synthesized phrases give noticeable improvement on the rhythm of the utterance. Although the change is not dramatic, it is certain that there is enhancement over other strategy for inter-syllable pause such as zero, fixed or fixed-ratio pauses. However, the application of these pause insertion rules only improves the perceptual naturalness slightly. Since the timing property is merely one of the many different prosodic features, it will not affect the overall perceived quality considerably.

Figure 7.9 illustrates the insertion of inter-syllable pause. It shows how the experimental measurement is applied to practical system.

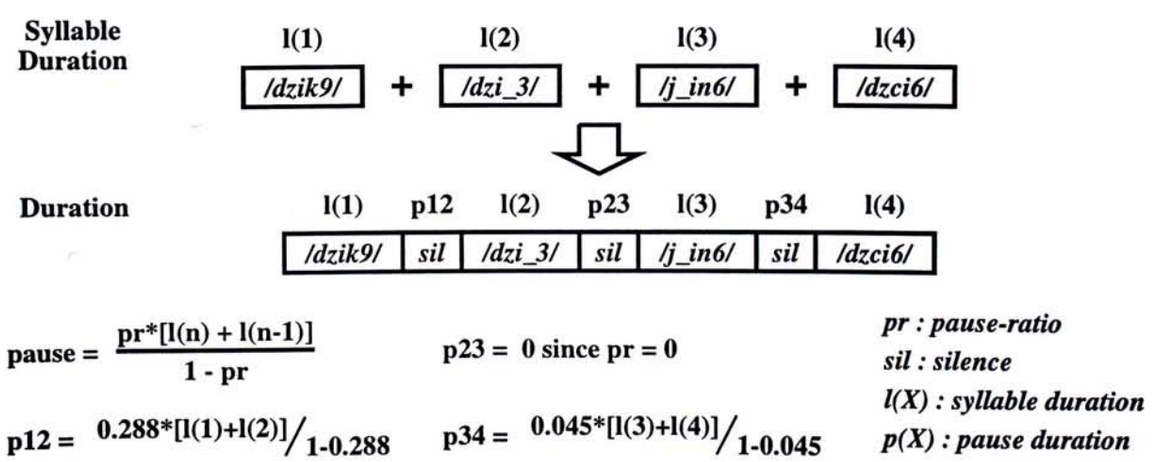

Figure 7.9: The inter-syllable pause duration is assigned to pauses between syllables in a phrase. It is based on the pause-ratio from the measurement and estimate the pause duration from the expected phrase duration.

From the comparison of the subjective quality of phrases with the pause inserted according to the experimental result and those with fixed pause or even none, slight improvement is noticed. Although the significance of pause duration is minor in the understanding of phrases, they play a significant role in the speech quality. Inter-syllable pause insertion is also useful in the speech synthesis by template concatenation. Properly assigned inter-syllable pause can improve the naturalness of the concatenated output.

7.5 Summary

To summarize, it can been concluded that inter-syllable pause duration exists in most multi-syllable phrases. It is determined by the involved phones in the phone-pairs. In continuous speech, the existence of pause depends on the pause-related phones, while the duration of pause also depends on the involved phones.

It has been found that the existence of pause ratio is consistent throughout the experiment. The pause-ratio factor in phrases of different syllable count shows that there is only minor changes in phrases of different syllable counts. As expected, the larger the syllable count, the smaller the pause-ratio. The normalized factor for two-syllable phrases to that of three/four phrases is around 1.2 indicating a 20% increase in the pause-ratio for the shorter phrases.

By insertion of proper length of pauses, it is found that there is some improvement on the perceived rhythm of the output phrases. The rhythm is found to be enhanced over that with zero, fixed or fixed pause-ratio pause. The result confirms the role of the proper inter-syllable pauses in the perceptual quality. However, the improvement cannot be regarded as substantial. It is because there are many other factors determining the overall quality.

Under certain occasions, acoustic coarticulatory effects are observed in the inter-syllable transition. These coarticulatory transitions can help experimentally verify the concept of voice correspondence for various phones.

Chapter 8

Conclusion and Further Work

8.1 Conclusion

There are many different important applications of speech synthesis. Some of them are essential while others may be just enhancement to existing applications. As the complexity of the applications of speech synthesis continue to increase, more sophisticated synthesis techniques and synthesis controls are desired. Articulatory control is believed to possess high potential to be the ultimate solution. This is because it is based on the control mechanism in real speech production. It is also capable of providing arbitrarily high controllability. In this work, inter-disciplinary approach based on articulatory phonetics, speech synthesis and network approximation is used for speech synthesis that possesses simple and efficient control.

The use of artificial neural network as the basic synthesizer architectures enables an implicit and non-parametric storage of phone templates as network parameters. This approach allows the retrieval of the spectral templates of phone units for reliable reproduction of the acoustic properties. In addition, the neural network approach also provides non-linear approximation for the articulatory input controls that have no corresponding training templates. Therefore, the neural network approach enables the construction of synthesizer networks with selectable output quality and degree of controllability through the provision of training templates, the choice of network architecture and also the selection of input control parameters used in the synthesizer network.

The articulatory control for speech synthesis allows simple and efficient control for the speech synthesis process. Adoption of the parameters in vowel quadrilaterals ensures

effective control for the synthesis of most vowels. The use of the articulatory synthesis control also allows extension of the control space for extra controllability.

The idea of voice correspondence, on the other hand, makes use of the articulatory phonetics knowledge for the enhancement and effective simplification of the synthesis control. Together with proper prosodic control, the voice correspondence enables the synthesis of nearly all Cantonese consonants and vowels to be done in a simple way with just a few control parameters. During synthesis, the articulatory based control parameters allows intuitive specification of the parameter trajectories that are only constrained by physical limitation of the involved articulators.

In addition, the voice correspondence gives useful transition targets for the articulatory control. These targets enables a better transition quality across phonetic segments in the synthesized speech output. Since the correspondence assignment is based on articulation positions, it ensures proper and effective transition targets. The correspondence also provides a useful basis for the simplification of the articulatory space according to the requirement by using the correspondence points for synthesis and adoption of parallel synthesizer network.

In continuous speech, pause duration is one of the many important prosodic features. To certain extent, it determines the rhythm of the spoken utterance and thus the perceived naturalness. The statistical experiment on the measurement of inter-syllable pause duration in Cantonese phrases gives valuable information on the characteristics of this feature. It provides useful guideline in designing controlling rules for this particular prosodic properties in short and fluently uttered speech. It can also be used as the basis for extension to determine pause properties in longer utterance.

All in all, this work emphasizes on the novel use of artificial neural network for acoustic speech synthesis under the control of articulatory parameters. It provides an integrated technique for speech synthesis that bridges the gap between copy concatenation and articulatory synthesis. By using the neural network speech synthesizer with the articulatory control in a small prototype, it is found that this approach has already achieved fair (4 in a 10 point scale) level of perceptual naturalness in the reproduced short speech segments. The quality of the vowel-to-vowel transitions in diphthongs can even achieve fairly good (6-7 in a 10 point scale) grading.

It is by no means an ideal method but it contributes to the speech synthesis area by making a starting point on the use of articulatory control for speech synthesis other than conventional articulatory synthesis. Together with the neural network approach, it provides an non-parametric synthesis technique with articulatory control. The full potential of the proposed methods is still to be explored and room for further improvement is available.

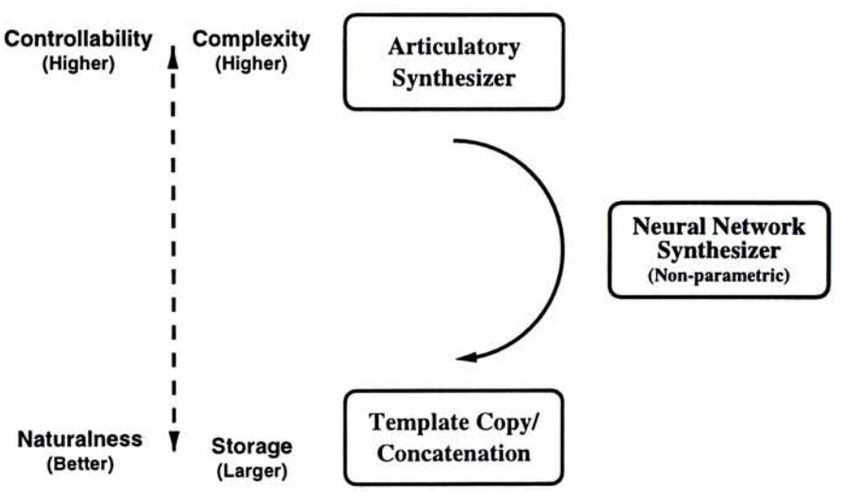

Figure 8.1: The neural network speech synthesis technique serves as an intermediate approach for bridging the gap between articulatory synthesis and copy concatenation using non-parametric templates.

8.2 Further Extension Work

Based on the fundamental work in this thesis, further extension could be continued to improve the performance of this approach. Emphasis of the suggested extensions are all aiming at providing better approximation and quality. Extension of the work to related research area is also proposed for better exploitation of the result and idea in this work.

8.2.1 Regularization Network Optimized on ISD

It has been found that synthesizer being trained by using a modified Itakura-Saito distortion (MISD) measure gives better perceptual quality. In addition, radial basis functions are popularly used for multi-variate approximation, it is employed as the synthesizer architecture in an attempt to test its capability in speech signal synthesis. Gaussian radial basis function is actually an optimized solution of sum square error with regularization on its second derivatives. Therefore, it is believed that using the MISD as the error criteria may be advantageous. Perhaps using the optimized analytical basis function as approximation basis in speech application will give better performance. The solution to the regularized approximation is surely highly complex which may hinder its wide application. However, it provides valuable information in the interpolation of spectral property in speech applications. Furthermore, the regularization criteria may also be changed to achieve desired properties in the optimized result.

8.2.2 Incorporation of Non-Articulatory Parameters to Control Space

Besides adding other control parameters from the articulatory space to the synthesizer network for enhancing controllability and quality, pitch and energy level may also be incorporated. The incorporation of pitch and/or energy as synthesizer control allows the spectral properties variation of phone units due to difference in pitch and energy level to be synthesized. Addition of parameters beyond the articulatory level enables other phonetic or acoustic knowledge to be accounted for by the synthesizer architecture. The adoption of abstract parameters as synthesis control is also an inherent characteristic of the neural network approach. Based on the training templates, network of proper size

will figure out the mapping of the control space to the domain of the output templates. These kinds of control may enable greater abstraction of the mapping between the control space and the output domain.

8.2.3 Experiment on Other Prosodic Features

Another possible extension is the collection of more data on prosodic features and to study their general characteristics. Better prosodic control on the synthesis process enable better perceptual naturalness on the synthesized output. Taking and analysis data to extract the general characteristics may seem trivial. However, thorough knowledge in these properties is undoubtedly beneficial to the production of more natural output speech.

8.2.4 Application of Voice Correspondence to Cantonese Coda Discrimination

The discrimination of Cantonese nasal and stop codas are found to be difficult because of their short duration. Based on the idea of voice correspondence (related to place of articulation), the synthesis of these codas are found to be successful with fairly good quality. Therefore, it is suggested to make use of the voice correspondence for assisting the discrimination of nasal and stop coda in Cantonese speech recognition. By finding out the correspondence target in the transition, it gives better confidence measure in discriminating the nasal and stop codas. Correspondence of nasal initials may also improve its discrimination. As a result, the accuracy of recognition of these consonants by conventional methods such as HMM, ANN, DTW etc. will be improved.

Appendix A

Cantonese Initials and Finals

A.1 Tables of All Cantonese Initials and Finals

Initial				
IPA	**Sydney Lau**	**CUHK-EE**	**Example**	**Remarks**
		nu	丫	Null initial
p	b	b_	巴	
t	d	d_	打	
k	g	g_	加	
k^w	gw	gw	瓜	
p^h	p	p_	趴	
t^h	t	t_	他	
k^h	k	k_	卡	
k^{wh}	kw	kw	誇	
s, ʃ	s	s_	沙	/ʃ/ before /y/
f	f	f_	花	
h	h	h_	蝦	
ts^h, $tʃ^h$	ch	ts	叉	/tʃ/ before /y/
ts, tʃ	j	dz	渣	/dʃ/ before /y/
l	l	l_	啦	
m	m	m_	媽	
n	n	n_	拿	
ŋ	ng	ng	啞	
w	w	w_	娃	
j	y	j_	也	

Table A.1: Table of initial consonants

Finals : Simple Vowels			
IPA	**Sydney Lau**	**CUHK-EE**	**Example**
a	aa	a_	沙
i	i	i_	司
u	u	u_	姑
ɛ	e	e_	些
ɔ	o	c_	梳
œ	oe	j_	朵
y	yu	y_	書

(a) Simple vowel finals

Finals : Diphthongs			
IPA	**Sydney Lau**	**CUHK-EE**	**Example**
ai	aai	ai	曬
au	aau	au	筲
ɐi	ai	ri	西
ɐu	au	ru	收
iu	iu	iu	消
ui	ui	ui	灰
ei	ei	ei	死
ci	oi	ci	鱠
ou	ou	cu	蘇
ɸy	eoi	jy	雖

(b) Diphthong finals

Finals : Nasal Sound Codas			
IPA	**Sydney Lau**	**CUHK-EE**	**Example**
/m/	m	m	吾
/ŋ/	ng	ng	五

(c) Nasal sound finals

Table A.2: Tables of finals without coda

Finals : Vowel-Nasals			
IPA	**Sydney Lau**	**CUHK-EE**	**Example**
am	aam	am	三
an	aan	an	山
ang	aang	a5	省
ɐm	am	rm	心
ɐn	an	rn	新
ɐŋ	ang	r5	牲
im	im	im	閃
in	in	in	先
ɪŋ	ing	i5	星
un	un	un	歡
ʊŋ	ung	u5	嵩
ɛŋ	eng	e5	腥
ɔn	on	cn	干
ɔŋ	ong	c5	桑
ɸn	eon	jn	詢
œŋ	oeng	j5	商
yn	yun	yn	孫

(a) Vowel-nasal finals

Finals : Vowel-Stop Codas			
IPA	**Sydney Lau**	**CUHK-EE**	**Example**
ap	aap	ap	圾
at	aat	at	殺
ak	aak	ak	索
ɐp	ap	rp	濕
ɐt	at	rt	失
ɐk	ak	rk	塞
ip	ip	ip	攝
it	it	it	屑
ɪk	ik	ik	色
ut	ut	ut	活
ʊk	uk	uk	肅
ɛk	ek	ek	石
ɔt	ot	ct	割
ɔk	ok	ck	角
ɸt	eot	jt	卒
œk	oek	jk	削
yt	yut	yt	說

(b) Vowel-stop finals

Table A.3: Tables of finals with coda

Appendix B

Using Distortion Measure as Error Function in Neural Network

B.1 Formulation of Itakura-Saito Distortion Measure for Neural Network Error Function

For a targeted spectrum t and estimated spectrum a, we define T and A as their log-spectrum respectively. The log-spectral difference, V is defined as

$$V = T - A \tag{B.1}$$

Itakura-Saito distortion measure is defined as

$$ISD = \frac{t}{a} - \log\frac{t}{a} - 1 \tag{B.2}$$

Substituting B.1 into B.2, we have

$$ISD = e^V - V - 1 \tag{B.3}$$

Formulating the ISD as neural network error function and applying error backpropagation for training, we define error function as, E as ISD.

$$E = e^V - V - 1 \tag{B.4}$$

In order to calculate the gradient of the error function with respect to the log-spectral estimation, E is differentiated with respect to A. Applying chain rule,

$$\frac{\delta E}{\delta A} = \frac{\delta E}{\delta V} \cdot \frac{\delta V}{\delta A} \tag{B.5}$$

and

$$\frac{\delta E}{\delta V} = \frac{\delta}{\delta V}(e^V - V - 1) \tag{B.6}$$

$$= e^V - 1 \tag{B.7}$$

On the other hand,

$$\frac{\delta V}{\delta A} = \frac{\delta(T - A)}{\delta A} \tag{B.8}$$

$$= -1 \tag{B.9}$$

Therefore, by substituting B.6 and B.8 into B.5, we get

$$-\frac{\delta E}{\delta A} = e^V - 1 \tag{B.10}$$

This formulation of the gradient of error surface shows that the emphasis is put on to the region with target values larger than the estimation. For regions of small targets or small difference, the emphasis is lowered. These are the key features of the ISD measure which preempt the spectral peaks in the target spectrum. In addition, due to the error gradient of the ISD measure, it is not suitable for use as error function in neural networks trained with error backpropagation technique.

B.2 Formulation of a Modified Itakura-Saito Distortion (MISD) Measure for Neural Network Error Function

Supposed that we have target spectrum t and estimated spectrum a, we define a new error measure as

$$E = \frac{1}{2}\sum_{\omega}\left(\frac{t}{a}-\frac{a}{t}+ln\frac{t}{a}\right)^2 \tag{B.11}$$

If log-spectrum is used instead of the magnitude spectrum, substitute $T = \log t$ and $A = \log a$, B.11 becomes

$$E = \frac{1}{2}\sum_{\omega}\left(e^{T-A}-e^{A-T}+T-A\right)^2 \tag{B.12}$$

Substituting $V = T - A$, B.12 becomes

$$E = \frac{1}{2}\sum_{\omega}\left(e^{V}-e^{-V}+V\right)^2 \tag{B.13}$$

Differentiating B.12 with respect to the estimation, A

$$\frac{\delta E}{\delta A} = \frac{\delta E}{\delta V}\cdot\frac{\delta V}{\delta A} \tag{B.14}$$

we have,

$$\begin{aligned}
\frac{\delta E}{\delta V} &= \left[e^{V}-e^{-V}+V\right]\left[e^{V}+e^{-V}+1\right] \\
&= e^{2V}+1+e^{V}-1-e^{-2V}-e^{-V}+Ve^{V}+Ve^{-V}+V \\
&= e^{2V}-e^{-2V}+e^{V}-e^{-V}+Ve^{V}+Ve^{-V}+V \\
&= e^{2V}+2Ve^{V}+V^2-Ve^{V} \\
&\quad -e^{-2V}+2Ve^{-V}-V^2-Ve^{-V} \\
&\quad +e^{V}-e^{-V}+V \\
&= \left[e^{V}+V\right]^2-\left[e^{-V}-V\right]^2+(1-V)e^{V}-(1+V)e^{-V}+V
\end{aligned} \tag{B.15}$$

Since $\frac{\delta V}{\delta A} = -1$, therefore

$$-\frac{\delta E}{\delta A} = V+\left[e^{V}+V\right]^2+(1-V)\,e^{V}-\left[e^{-V}-V\right]^2-(1+V)\,e^{-V} \tag{B.16}$$

which can be plugged into the error backporpagation algorithm for network training using the modified Itakura-Saito distortion.

Appendix C

Orthogonal Least Square Learning Algorithm for Raidal Basis Function Network Training

C.1 Orthogonal Least Squares Learning Algorithm for Radial Basis Function Network Training

The algorithm is based on that of Chen et. al [104] for the training of radial basis function (RBF) approximation network.

The problem of radial basis function approximation is given in general as

$$f(x) = \lambda_0 + \sum_{i=1}^{M} \lambda_i \phi(\|x - c_i\|) \tag{C.1}$$

Suppose that we have N pairs of targeted outputs d_i with inputs x_i, for $i = 1, \cdots N$, the problem is formulated as a regression problem.

$$d(i) = \sum_{j=1}^{M} p_j(i)\theta_j + \epsilon(i) \tag{C.2}$$

where $p_j(i)$ are regressors which are functions of inout $x_j(i)$. The error, $\epsilon(i)$, is assumed to be uncorrelated with the regressors, $p_j(i)$. For fixed center c_j with non-linearity, $\phi(\cdot)$, it corresponds to a regressor, $p_j(i)$. The center selection for the radial basis function from training templates is regarded as subset selection of regressors from candidate set. By reformulating the problem in matrix form, we have

$$d = P\Theta + E \tag{C.3}$$

with

$$d = [d(1) \cdots d(N)]^t \tag{C.4}$$

$$E = [\epsilon(1) \cdots \epsilon(N)]^t \tag{C.5}$$

$$\Theta = [\theta(1) \cdots \theta(N)]^t \tag{C.6}$$

$$P = \begin{bmatrix} p_1(1) & \cdots & p_M(1) \\ \vdots & & \vdots \\ p_1(N) & \cdots & p_M(N) \end{bmatrix} \tag{C.7}$$

where P is made up of rows of regressors from the training data.

Because the regressors are correlated in general, the OLS scheme will first form an uncorrelated basis matrix, W such that

$$P = WA \tag{C.8}$$

where W is NxM and A is MxM in size. Moreover, A has a diagonal of '1's with all elements below the diagonal equal '0'. W is now a set of orthogonal columns w_i such that

$$W^tW = H \tag{C.9}$$

where H is a diagonal matrix.

Therefore, the problem is now formulated as

$$d = Wg + E \tag{C.10}$$

The estimated parameters, $\tilde{g}$ is given by

$$\tilde{g} = H^{-1}W^td \tag{C.11}$$

and the final estimated solution for the parameters, $\tilde{\Theta}$ is

$$\tilde{\Theta} = A^{-1}\tilde{g} \tag{C.12}$$

Appendix D

Phrase Lists

D.1 Two-Syllable Phrase List for the Pause Duration Experiment

D.1.1 兩字詞

/w_ai/ — /j_rn/	1	○	壞人
/d_ai/ — /s_i_/	2	○	大事
/t_ri/ — /t_ip/	3	○	體貼
/tsai/ — /d_ck/	4	○	猜度
/d_ei/ — /tsan/	5	○	地產
/l_ci/ — /n_in/	6	○	來年
/g_ai/ — /dzi_/	7	○	戒指
/s_e_/ — /j_i_/	8	○	寫意
/tse_/ — /s_i_/	9	○	車匙
/dze_/ — /t_in/	10	○	蔗田
/s_e_/ — /d_am/	11	○	蛇膽
/tse_/ — /tsi_/	12	○	奢侈
/tse_/ — /n_ci/	13	○	車內
/j_e_/ — /dzi_/	14	○	椰子
/d_jy/ — /j_in/	15	○	兌現
/s_jy/ — /s_ru/	16	○	稅收
/j_y_/ — /t_ru/	17	○	魚頭
/s_y_/ — /d_im/	18	○	書店
/j_y_/ — /tsi_/	19	○	魚翅
/j_y_/ — /n_ri/	20	○	淤泥
/dzy_/ — /dzik/	21	○	主席
/j_i_/ — /j_yt/	22	○	二月
/j_i_/ — /s_i_/	23	○	依時
/j_i_/ — /t_cu/	24	○	意圖
/j_i_/ — /d_ei/	25	○	異地
/j_i_/ — /tsin/	26	○	以前
/j_i_/ — /n_an/	27	○	疑難
/j_i_/ — /dzi_/	28	○	兒子
/s_ru/ — /j_rp/	29	○	收入
/l_ru/ — /s_rm/	30	○	留心
/k_ru/ — /t_ru/	31	○	叩頭
/g_ru/ — /d_ei/	32	○	舊地
/tsru/ — /tsim/	33	○	抽籤
/j_ru/ — /n_ei/	34	○	油膩
/tsru/ — /dzj5/	35	○	抽獎
/s_a_/ — /j_ru/	36	○	沙丘
/b_a_/ — /s_i_/	37	○	巴士
/m_a_/ — /t_ri/	38	○	馬蹄
/w_a_/ — /d_u5/	39	○	華東
/g_a_/ — /tsit/	40	○	假設
/g_a_/ — /n_jy/	41	○	嫁女
/gwa_/ — /dzi_/	42	○	瓜子
/d_c_/ — /j_rp/	43	○	墮入
/d_c_/ — /s_cu/	44	○	多數
/f_c_/ — /t_c5/	45	○	課堂
/ngc_/ — /d_ri/	46	○	臥底
/f_c_/ — /tse_/	47	○	貨車
/d_c_/ — /n_in/	48	○	多年

/dzc_/ — /dzi_/	49	○	阻止	/f_c_/ — /b_rn/	84	○	貨品
/w_u_/ — /j_i5/	50	○	弧形	/w_u_/ — /p_un/	85	○	烏盆
/w_u_/ — /s_i_/	51	○	護士	/w_u_/ — /w_ri/	86	○	污濊
/w_u_/ — /t_ru/	52	○	芋頭	/w_u_/ — /m_in/	87	○	互勉
/w_u_/ — /d_u5/	53	○	烏冬	/w_u_/ — /b_un/	88	○	湖畔
/w_u_/ — /tsak/	54	○	户册	/s_ri/ — /f_c5/	89	○	西方
/w_u_/ — /n_am/	55	○	湖南	/f_ui/ — /h_rn/	90	○	悔恨
/w_u_/ — /dzik/	56	○	污漬	/s_ri/ — /l_cu/	91	○	細路
/t_ai/ — /p_i5/	57	○	太平	/d_ai/ — /g_a_/	92	○	大家
/g_ai/ — /w_a_/	58	○	佳話	/tsai/ — /k_yn/	93	○	猜拳
/k_ri/ — /m_u5/	59	○	啓蒙	/s_ri/ — /ngru/	94	○	犀牛
/d_ri/ — /b_cu/	60	○	逮捕	/j_e_/ — /f_c_/	95	○	野火
/tse_/ — /p_ai/	61	○	車牌	/s_e_/ — /h_j5/	96	○	麝香
/tse_/ — /w_ri/	62	○	車位	/dze_/ — /l_ri/	97	○	謝禮
/j_e_/ — /m_an/	63	○	夜晚	/dze_/ — /g_jy/	98	○	借據
/s_e_/ — /b_a_/	64	○	射靶	/tse_/ — /k_a_/	99	○	車卡
/s_jy/ — /p_in/	65	○	碎片	/j_e_/ — /ngru/	100	○	野牛
/s_y_/ — /w_un/	66	○	舒緩	/s_y_/ — /f_c5/	101	○	書房
/dzjy/ — /m_i5/	67	○	罪名	/dzjy/ — /h_ru/	102	○	最後
/d_jy/ — /b_ei/	68	○	對比	/tsjy/ — /l_uk/	103	○	翠綠
/j_i_/ — /p_i5/	69	○	夷平	/dzjy/ — /g_ru/	104	○	追究
/j_i_/ — /w_ak/	70	○	疑惑	/d_jy/ — /k_c5/	105	○	對抗
/j_i_/ — /m_ei/	71	○	意味	/tsy_/ — /ngri/	106	○	廚藝
/j_i_/ — /b_c5/	72	○	異邦	/j_i_/ — /f_uk/	107	○	衣服
/k_iu/ — /p_ai/	73	○	橋牌	/j_i_/ — /h_ru/	108	○	以後
/j_ru/ — /w_i5/	74	○	游泳	/j_i_/ — /l_e5/	109	○	衣領
/j_ru/ — /m_rk/	75	○	油墨	/j_i_/ — /g_in/	110	○	意見
/tsru/ — /b_ei/	76	○	籌備	/j_i_/ — /k_au/	111	○	依靠
/f_a_/ — /p_i5/	77	○	花瓶	/j_i_/ — /ngci/	112	○	意外
/f_a_/ — /w_ri/	78	○	花卉	/s_iu/ — /f_a_/	113	○	消化
/s_a_/ — /m_ck/	79	○	沙漠	/tsru/ — /h_u5/	114	○	抽空
/m_a_/ — /b_cu/	80	○	麻布	/t_ru/ — /l_an/	115	○	偷懶
/p_c_/ — /p_c_/	81	○	婆婆	/dziu/ — /g_rp/	116	○	焦急
/f_c_/ — /w_an/	82	○	科幻	/s_ru/ — /kwri/	117	○	羞愧
/g_c_/ — /m_ri/	83	○	歌迷	/j_ru/ — /nga_/	118	○	優雅

/g_a_/ — /f_at/	119	○	家法
/m_a_/ — /h_ei/	120	○	馬戲
/m_a_/ — /l_cu/	121	○	馬路
/f_a_/ — /g_ap/	122	○	花甲
/f_a_/ — /k_ru/	123	○	花球
/m_a_/ — /ngri/	124	○	螞蟻
/h_c_/ — /f_c5/	125	○	何方
/f_c_/ — /h_ck/	126	○	科學
/h_c_/ — /l_in/	127	○	可憐
/f_c_/ — /g_jy/	128	○	科舉
/g_c_/ — /k_ek/	129	○	歌劇
/g_c_/ — /ngri/	130	○	歌藝
/w_u_/ — /f_c5/	131	○	互訪
/w_u_/ — /h_ru/	132	○	餬口
/w_u_/ — /l_ei/	133	○	狐狸
/w_u_/ — /g_a_/	134	○	護駕
/w_u_/ — /k_jy/	135	○	污渠
/w_u_/ — /nga_/	136	○	烏鴉
/l_in/ — /j_rm/	137	○	連任
/d_an/ — /s_rn/	138	○	單身
/s_in/ — /t_in/	139	○	先天
/n_in/ — /d_ri/	140	○	年底
/g_in/ — /tsi_/	141	○	堅持
/nucn/ — /n_i5/	142	○	安寧
/j_in/ — /dzuk/	143	○	延續
/s_j5/ — /j_jk/	144	○	商約
/dzu5/ — /s_rm/	145	○	忠心
/l_i5/ — /t_i5/	146	○	聆聽
/tsj5/ — /d_cu/	147	○	倡導
/j_i5/ — /tsru/	148	○	應酬
/m_i5/ — /n_in/	149	○	明年
/j_i5/ — /dzrn/	150	○	認眞
/j_rm/ — /j_ru/	151	○	任由
/tsrm/ — /s_i_/	152	○	沉思
/s_rm/ — /t_ai/	153	○	心態
/s_rm/ — /d_cu/	154	○	深度
/t_am/ — /tsat/	155	○	探測
/g_rm/ — /n_in/	156	○	今年
/tsrm/ — /dzim/	157	○	侵佔
/tsap/ — /j_rp/	158	○	插入
/dzap/ — /s_an/	159	○	集散
/h_rp/ — /t_u5/	160	○	合同
/tsrp/ — /d_uk/	161	○	緝毒
/d_ap/ — /tsjt/	162	○	踏出
/s_rp/ — /n_in/	163	○	十年
/s_rp/ — /dzuk/	164	○	十足
/tsrt/ — /j_yt/	165	○	七月
/s_rt/ — /s_ru/	166	○	失手
/tsjt/ — /t_i5/	167	○	出庭
/b_it/ — /d_i5/	168	○	必定
/tsjt/ — /tsai/	169	○	出差
/tsjt/ — /n_ap/	170	○	出納
/d_it/ — /dzjy/	171	○	秩序
/s_ik/ — /j_i5/	172	○	適應
/l_ik/ — /s_i_/	173	○	歷時
/s_ek/ — /t_ru/	174	○	石頭
/g_ik/ — /d_yn/	175	○	極端
/d_uk/ — /tsuk/	176	○	督促
/g_ik/ — /n_an/	177	○	極難
/d_rk/ — /dzi5/	178	○	特徵
/t_in/ — /p_i5/	179	○	填平
/l_in/ — /w_an/	180	○	連環
/tsin/ — /m_in/	181	○	前面
/s_an/ — /b_an/	182	○	舢舨
/tsu5/ — /p_un/	183	○	重判
/j_u5/ — /w_ui/	184	○	融匯
/l_u5/ — /m_au/	185	○	龍貓
/p_a5/ — /b_ai/	186	○	澎湃
/g_rm/ — /p_cu/	187	○	金舖
/j_rm/ — /w_rn/	188	○	音韻

/tsrm/ — /m_rk/	189	◯	沉默
/j_rm/ — /b_ei/	190	◯	蔭庇
/tsrp/ — /p_c_/	191	◯	緝破
/dzap/ — /w_ui/	192	◯	雜匯
/dzap/ — /m_rt/	193	◯	雜物
/h_rp/ — /b_ik/	194	◯	合璧
/t_it/ — /p_a5/	195	◯	鐵棚
/b_rt/ — /w_ak/	196	◯	筆劃
/s_rt/ — /m_i5/	197	◯	失明
/d_yt/ — /b_iu/	198	◯	奪標
/j_uk/ — /p_ui/	199	◯	玉佩
/f_uk/ — /w_ut/	200	◯	復活
/d_jk/ — /m_c_/	201	◯	琢磨
/dzak/ — /b_ei/	202	◯	責備
/l_in/ — /f_a_/	203	◯	蓮花
/t_in/ — /h_ci/	204	◯	塡海
/n_in/ — /l_ik/	205	◯	年曆
/s_jn/ — /g_an/	206	◯	瞬間
/f_an/ — /k_c5/	207	◯	反抗
/h_in/ — /ngri/	208	◯	獻藝
/gwc5/ — /f_uk/	209	◯	光復
/tsc5/ — /h_ci/	210	◯	滄海
/h_j5/ — /l_ei/	211	◯	鄉里
/s_j5/ — /g_a_/	212	◯	商家
/tsj5/ — /k_ru/	213	◯	搶購
/l_c5/ — /nga_/	214	◯	狼牙
/s_rm/ — /f_a_/	215	◯	深化
/h_rm/ — /h_ci/	216	◯	陷害
/g_am/ — /l_am/	217	◯	橄欖
/g_am/ — /g_ai/	218	◯	尷尬
/g_rm/ — /k_c5/	219	◯	金礦
/n_am/ — /nga_/	220	◯	南亞
/s_rp/ — /f_c5/	221	◯	拾荒
/dzap/ — /h_rp/	222	◯	集合
/h_ip/ — /l_ik/	223	◯	協力
/h_rp/ — /g_ak/	224	◯	合格
/s_ip/ — /k_rp/	225	◯	涉及
/h_ap/ — /ngai/	226	◯	狹隘
/m_it/ — /f_c_/	227	◯	滅火
/tsit/ — /h_rp/	228	◯	切合
/j_it/ — /l_au/	229	◯	熱鬧
/dzit/ — /g_i5/	230	◯	捷徑
/l_it/ — /k_j5/	231	◯	列強
/t_it/ — /ngru/	232	◯	鐵勾
/l_ck/ — /f_c_/	233	◯	落貨
/dzik/ — /h_i5/	234	◯	即興
/g_ik/ — /l_ck/	235	◯	極樂
/t_ck/ — /g_un/	236	◯	托管
/g_ik/ — /k_ru/	237	◯	擊球
/s_ik/ — /ngan/	238	◯	食晏

D.2 Three/Four-Syllable Phrase List for the Pause Duration Experiment

D.2.1 片語

/w_ai/ – /j_rn/ – /d_c5/ – /d_cu/	1	○	壞人當道
/gwck/ – /g_a_/ – /d_ai/ – /s_i_/	2	○	國家大事
/t_ri/ – /t_ip/ – /j_rp/ – /m_ei/	3	○	體貼入微
/s_jn/ – /s_jy/ – /tsai/ – /d_ck/	4	○	純粹猜度
/d_ei/ – /tsan/ – /g_u5/ – /s_i_/	5	○	地產公司
/l_ci/ – /n_in/ – /h_cu/ – /w_rn/	6	○	來年好運
/dzyn/ – /s_ek/ – /g_ai/ – /dzi_/	7	○	鑽石戒指
/s_e_/ – /j_i_/ – /s_y_/ – /s_ik/	8	○	寫意舒適
/h_ru/ – /b_ei/ – /tse_/ – /s_i_/	9	○	後備車匙
/g_rm/ – /dze_/ – /t_in/ –	10	○	甘蔗田
/d_uk/ – /s_e_/ – /d_am/ –	11	○	毒蛇膽
/tse_/ – /tsi_/ – /b_rn/ –	12	○	奢侈品
/f_c_/ – /tse_/ – /n_ci/ –	13	○	火車內
/j_e_/ – /dzi_/ – /s_y_/ –	14	○	椰子樹
/dzi_/ – /p_iu/ – /d_jy/ – /j_in/	15	○	支票兌現
/dzr5/ – /g_a_/ – /s_jy/ – /s_ru/	16	○	增加稅收
/j_y_/ – /t_ru/ – /d_ru/ – /f_u_/	17	○	魚頭豆腐
/s_rn/ – /w_a_/ – /s_y_/ – /d_im/	18	○	新華書店
/s_j5/ – /d_r5/ – /j_y_/ – /tsi_/	19	○	上等魚翅
/j_y_/ – /n_ri/ – /dzc_/ – /s_rk/	20	○	淤泥阻塞
/gwck/ – /g_a_/ – /dzy_/ – /dzik/	21	○	國家主席
/j_i_/ – /j_yt/ – /s_rp/ – /j_rt/	22	○	二月十日
/j_i_/ – /s_i_/ – /j_i_/ – /h_ru/	23	○	依時依候
/j_i_/ – /t_cu/ – /b_rt/ – /gwri/	24	○	意圖不軌
/s_rn/ – /tsy_/ – /j_i_/ – /d_ei/	25	○	身處異地
/h_rn/ – /g_ru/ – /j_i_/ – /tsin/	26	○	很久以前
/g_ai/ – /k_yt/ – /j_i_/ – /n_an/	27	○	解決疑難
/tsrn/ – /s_r5/ – /j_i_/ – /dzi_/	28	○	親生兒子
/g_i5/ – /dzri/ – /s_ru/ – /j_rp/	29	○	經濟收入
/l_ru/ – /s_rm/ – /s_j5/ – /f_c_/	30	○	留心上課

/k_ru/ – /t_ru/ – /dze_/ – /l_ri/	31	◯	叩頭謝禮
/g_ru/ – /d_ei/ – /tsu5/ – /j_ru/	32	◯	舊地重遊
/tsru/ – /tsim/ – /k_yt/ – /d_i5/	33	◯	抽籤決定
/j_ru/ – /n_ei/ – /s_ik/ – /m_rt/	34	◯	油膩食物
/d_ai/ – /tsru/ – /dzj5/ –	35	◯	大抽獎
/s_a_/ – /j_ru/ – /dzi_/ – /s_j5/	36	◯	沙丘之上
/s_jy/ – /d_cu/ – /b_a_/ – /s_i_/	37	◯	隧道巴士
/m_a_/ – /t_ri/ – /t_it/ –	38	◯	馬蹄鐵
/w_a_/ – /d_u5/ – /s_jy/ – /dzci/	39	◯	華東水災
/g_a_/ – /tsit/ – /tsc_/ – /nung/	40	◯	假設錯誤
/g_a_/ – /n_jy/ – /b_e5/ –	41	◯	嫁女餅
/d_ai/ – /h_u5/ – /gwa_/ – /dzi_/	42	◯	大紅瓜子
/d_c_/ – /j_rp/ – /ngci/ – /h_c_/	43	◯	墮入愛河
/d_c_/ – /s_cu/ – /j_rn/ – /s_i_/	44	◯	多數人士
/f_c_/ – /t_c5/ – /dzi_/ – /s_j5/	45	◯	課堂之上
/ngc_/ – /d_ri/ – /t_am/ – /j_yn/	46	◯	臥底探員
/f_c_/ – /tse_/ – /s_i_/ – /g_ei/	47	◯	貨車司機
/d_c_/ – /n_in/ – /g_i5/ – /j_im/	48	◯	多年經驗
/k_rp/ – /s_i_/ – /dzc_/ – /dzi_/	49	◯	及時阻止
/w_u_/ – /j_i5/ – /t_cu/ – /nucn/	50	◯	弧形圖案
/w_u_/ – /s_i_/ – /h_ck/ – /h_au/	51	◯	護士學校
/j_e_/ – /s_r5/ – /w_u_/ – /t_ru/	52	◯	野生芋頭
/w_u_/ – /d_u5/ – /m_in/ –	53	◯	烏冬麵
/tsyn/ – /dzrn/ – /w_u_/ – /tsat/	54	◯	村鎮户册
/w_u_/ – /n_am/ – /s_a5/ –	55	◯	湖南省
/tsi5/ – /tsjy/ – /w_u_/ – /dzik/	56	◯	清除污漬
/t_ai/ – /p_i5/ – /s_i5/ – /s_ri/	57	◯	太平盛世
/tsyn/ – /w_ri/ – /g_ai/ – /w_a_/	58	◯	傳爲佳話
/k_ri/ – /m_u5/ – /tsin/ – /b_ui/	59	◯	啓蒙前輩
/d_ri/ – /b_cu/ – /j_i_/ – /f_an/	60	◯	逮捕疑犯
/tse_/ – /p_ai/ – /h_cu/ – /m_a_/	61	◯	車牌號碼
/s_i_/ – /dzcu/ – /tse_/ – /w_ri/	62	◯	時租車位
/dzik/ – /dzi_/ – /j_e_/ – /m_an/	63	◯	直至夜晚
/s_e_/ – /b_a_/ – /l_in/ – /dzap/	64	◯	射靶練習
/b_c_/ – /l_ei/ – /s_jy/ – /p_in/	65	◯	玻璃碎片

/s_y_/ – /w_un/ – /g_rn/ – /dzj5/	66	○	舒緩緊張
/dzjy/ – /m_i5/ – /s_i5/ – /l_ap/	67	○	罪名成立
/w_u_/ – /s_j5/ – /d_jy/ – /b_ei/	68	○	互相對比
/j_i_/ – /p_i5/ – /d_ik/ – /gwck/	69	○	夷平敵國
/s_rm/ – /g_rm/ – /j_i_/ – /w_ak/	70	○	深感疑惑
/n_ci/ – /n_jy/ – /j_i_/ – /m_ei/	71	○	內裡意味
/t_ru/ – /b_rn/ – /j_i_/ – /b_c5/	72	○	投奔異邦
/k_iu/ – /p_ai/ – /b_ei/ – /tsci/	73	○	橋牌比賽
/j_ru/ – /w_i5/ – /g_in/ – /dzj5/	74	○	游泳健將
/j_ru/ – /m_rk/ – /j_rn/ – /tsat/	75	○	油墨印刷
/tsru/ – /b_ei/ – /dzit/ – /m_uk/	76	○	籌備節目
/s_jy/ – /dzi5/ – /f_a_/ – /p_i5/	77	○	水晶花瓶
/f_a_/ – /w_ri/ – /dzin/ – /l_am/	78	○	花卉展覽
/gwc_/ – /b_ik/ – /s_a_/ – /m_ck/	79	○	戈壁沙漠
/tscu/ – /j_i_/ – /m_a_/ – /b_cu/	80	○	粗衣麻布
/p_c_/ – /p_c_/ – /m_a_/ – /m_a_/	81	○	婆婆媽媽
/f_c_/ – /w_an/ – /s_iu/ – /s_yt/	82	○	科幻小說
/j_it/ – /tsi5/ – /g_c_/ – /m_ri/	83	○	熱情歌迷
/tsrn/ – /l_it/ – /f_c_/ – /b_rn/	84	○	陳列貨品
/j_e_/ – /s_rm/ – /w_u_/ – /p_un/	85	○	夜審烏盆
/w_u_/ – /w_ri/ – /s_i_/ – /s_j5/	86	○	污濊思想
/w_u_/ – /l_ri/ – /w_u_/ – /m_in/	87	○	互勵互勉
/d_ai/ – /m_i5/ – /w_u_/ – /b_un/	88	○	大明湖畔
/s_ri/ – /f_c5/ – /dzit/ – /h_ck/	89	○	西方哲學
/f_ui/ – /h_rn/ – /d_c5/ – /n_in/	90	○	悔恨當年
/s_ri/ – /l_cu/ – /dzri/ –	91	○	細路仔
/d_ai/ – /g_a_/ – /h_rp/ – /dzck/	92	○	大家合作
/tsai/ – /k_yn/ – /j_ru/ – /h_ei/	93	○	猜拳遊戲
/b_ak/ – /s_ri/ – /ngru/ –	94	○	白犀牛
/j_e_/ – /f_c_/ – /w_ui/ –	95	○	野火會
/s_e_/ – /h_j5/ – /h_ei/ – /m_ei/	96	○	麝香氣味
/d_ap/ – /dze_/ – /l_ri/ –	97	○	答謝禮
/t_ai/ – /f_un/ – /dze_/ – /g_jy/	98	○	貸款借據
/f_c_/ – /tse_/ – /k_a_/ –	99	○	火車卡
/s_ri/ – /b_cu/ – /j_e_/ – /ngru/	100	○	西部野牛

/j_y_/ – /s_y_/ – /f_c5/ –	101	○	御書房
/dzjy/ – /h_ru/ – /t_u5/ – /d_ip/	102	○	最後通諜
/tsjy/ – /l_uk/ – /s_y_/ – /l_rm/	103	○	翠綠樹林
/dzjy/ – /g_ru/ – /dzat/ – /j_rm/	104	○	追究責任
/w_u_/ – /s_j5/ – /d_jy/ – /k_c5/	105	○	互相對抗
/tsy_/ – /ngri/ – /dzi5/ – /dzrm/	106	○	廚藝精湛
/j_i_/ – /f_uk/ – /h_ai/ – /m_rt/	107	○	衣服鞋襪
/dzi_/ – /tsi_/ – /j_i_/ – /h_ru/	108	○	自此以後
/tsi5/ – /g_it/ – /j_i_/ – /l_e5/	109	○	清潔衣領
/f_at/ – /b_iu/ – /j_i_/ – /g_in/	110	○	發表意見
/w_u_/ – /s_j5/ – /j_i_/ – /k_au/	111	○	互相依靠
/g_au/ – /t_u5/ – /j_i_/ – /ngci/	112	○	交通意外
/s_iu/ – /f_a_/ – /h_ri/ – /t_u5/	113	○	消化系統
/tsru/ – /h_u5/ – /tsjt/ – /dzik/	114	○	抽空出席
/dzik/ – /g_u_/ – /t_ru/ – /l_an/	115	○	藉故偷懶
/s_rm/ – /tsi5/ – /dziu/ – /g_rp/	116	○	心情焦急
/m_ei/ – /g_rm/ – /s_ru/ – /kwri/	117	○	未感羞愧
/g_ak/ – /d_iu/ – /j_ru/ – /nga_/	118	○	格調優雅
/g_a_/ – /f_at/ – /s_i_/ – /h_ru/	119	○	家法侍候
/m_a_/ – /h_ei/ – /t_yn/ –	120	○	馬戲團
/w_a5/ – /gwc_/ – /m_a_/ – /l_cu/	121	○	橫過馬路
/f_a_/ – /g_ap/ – /dzi_/ – /n_in/	122	○	花甲之年
/s_ru/ – /f_a_/ – /k_ru/ –	123	○	繡花球
/m_a_/ – /ngri/ – /s_j5/ – /s_y_/	124	○	螞蟻上樹
/h_c_/ – /f_c5/ – /s_rn/ – /s_i5/	125	○	何方神聖
/f_c_/ – /h_ck/ – /t_ai/ – /d_cu/	126	○	科學態度
/s_rn/ – /s_ri/ – /h_c_/ – /l_in/	127	○	身世可憐
/f_c_/ – /g_jy/ – /h_au/ – /s_i_/	128	○	科舉考試
/g_c_/ – /k_ek/ – /dzit/ – /m_uk/	129	○	歌劇節目
/g_c_/ – /ngri/ – /dzi5/ – /dzam/	130	○	歌藝精湛
/l_j5/ – /d_ei/ – /w_u_/ – /f_c5/	131	○	兩地互訪
/b_rt/ – /dzuk/ – /w_u_/ – /h_ru/	132	○	不足餬口
/w_u_/ – /l_ei/ – /dzi5/ –	133	○	狐狸精
/w_u_/ – /g_a_/ – /j_ru/ – /g_u5/	134	○	護駕有功
/tsi5/ – /l_ei/ – /w_u_/ – /k_jy/	135	○	清理污渠

/w_u_/ – /nga_/ – /kwrn/ –	136	◯	烏鴉群
/dzci/ – /d_cu/ – /l_in/ – /j_rm/	137	◯	再度連任
/d_an/ – /s_rn/ – /h_cn/ –	138	◯	單身漢
/s_in/ – /t_in/ – /k_yt/ – /h_rm/	139	◯	先天缺陷
/l_u5/ – /l_ik/ – /n_in/ – /d_ri/	140	◯	農曆年底
/g_ri/ – /dzuk/ – /g_in/ – /tsi_/	141	◯	繼續堅持
/nucn/ – /n_i5/ – /s_r5/ – /w_ut/	142	◯	安寧生活
/j_in/ – /dzuk/ – /s_r5/ – /m_i5/	143	◯	延續生命
/s_j5/ – /j_jk/ – /w_ui/ – /m_in/	144	◯	商約會面
/dzu5/ – /s_rm/ – /g_r5/ – /g_r5/	145	◯	忠心耿耿
/l_ru/ – /s_rm/ – /l_i5/ – /t_i5/	146	◯	留心聆聽
/tsj5/ – /d_cu/ – /g_ci/ – /g_ap/	147	◯	倡導改革
/j_i5/ – /tsru/ – /b_rn/ – /h_at/	148	◯	應酬賓客
/m_i5/ – /n_in/ – /n_in/ – /tsc_/	149	◯	明年年初
/t_ai/ – /d_cu/ – /j_i5/ – /dzrn/	150	◯	態度認眞
/j_rm/ – /j_ru/ – /tsy_/ – /dzi_/	151	◯	任由處置
/tsrm/ – /s_i_/ – /m_i5/ – /s_j5/	152	◯	沉思冥想
/s_rm/ – /t_ai/ – /b_rt/ – /l_cu/	153	◯	心態畢露
/h_ci/ – /s_jy/ – /s_rm/ – /d_cu/	154	◯	海水深度
/t_am/ – /tsat/ – /w_an/ – /g_i5/	155	◯	探測環境
/g_rm/ – /n_in/ – /n_in/ – /m_ei/	156	◯	今年年尾
/tsrm/ – /dzim/ – /t_cu/ – /d_ei/	157	◯	侵佔土地
/tsap/ – /j_rp/ – /dzi5/ – /g_in/	158	◯	插入證件
/dzap/ – /s_an/ – /d_ei/ –	159	◯	集散地
/h_rp/ – /t_u5/ – /f_at/ –	160	◯	合同法
/tsrp/ – /d_uk/ – /b_cu/ – /m_un/	161	◯	緝毒部門
/d_ap/ – /tsjt/ – /d_ai/ – /t_c5/	162	◯	踏出大堂
/s_rp/ – /n_in/ – /s_i_/ – /g_an/	163	◯	十年時間
/s_rp/ – /dzuk/ – /s_jn/ – /s_rm/	164	◯	十足信心
/tsrt/ – /j_yt/ – /s_rp/ – /s_ei/	165	◯	七月十四
/b_rt/ – /s_rn/ – /s_rt/ – /s_ru/	166	◯	不慎失手
/tsjt/ – /t_i5/ – /dzck/ – /dzi5/	167	◯	出庭作證
/b_it/ – /d_i5/ – /s_i5/ – /g_u5/	168	◯	必定成功
/tsjt/ – /tsai/ – /gwck/ – /ngci/	169	◯	出差國外
/tsjt/ – /n_ap/ – /gwri/ – /t_ci/	170	◯	出納櫃檯

/w_ri/ – /tsi_/ – /d_it/ – /dzjy/	171	◯	維持秩序
/s_ik/ – /j_i5/ – /b_in/ – /f_a_/	172	◯	適應變化
/l_ik/ – /s_i_/ – /h_rn/ – /g_ru/	173	◯	歷時很久
/s_ek/ – /t_ru/ – /d_iu/ – /h_rk/	174	◯	石頭雕刻
/g_ik/ – /d_yn/ – /dzy_/ – /j_i_/	175	◯	極端主義
/d_uk/ – /tsuk/ – /g_u5/ – /dzck/	176	◯	督促工作
/g_ik/ – /n_an/ – /s_i5/ – /g_u5/	177	◯	極難成功
/s_j5/ – /m_au/ – /d_rk/ – /dzi5/	178	◯	相貌特徵
/t_in/ – /p_i5/ – /h_ci/ – /g_c5/	179	◯	塡平海港
/l_in/ – /w_an/ – /f_at/ – /s_e_/	180	◯	連環發射
/tsin/ – /m_in/ – /s_ru/ – /l_cu/	181	◯	前面修路
/s_an/ – /b_an/ – /s_iu/ – /t_e5/	182	◯	舢舨小艇
/tsu5/ – /p_un/ – /s_ei/ – /j_i5/	183	◯	重判死刑
/j_u5/ – /w_ui/ – /g_un/ – /t_u5/	184	◯	融匯貫通
/l_u5/ – /m_au/ – /k_a_/ – /t_u5/	185	◯	龍貓卡通
/l_c5/ – /t_cu/ – /p_a5/ – /b_ai/	186	◯	浪濤澎湃
/g_rm/ – /p_cu/ – /ngrn/ – /h_cu/	187	◯	金舖銀號
/j_rm/ – /w_rn/ – /j_yt/ – /j_i_/	188	◯	音韻悅耳
/tsrm/ – /m_rk/ – /b_rt/ – /j_in/	189	◯	沉默不言
/dzcu/ – /s_in/ – /j_rm/ – /b_ei/	190	◯	祖先蔭庇
/tsrp/ – /p_c_/ – /tsat/ – /tsau/	191	◯	緝破賊巢
/d_ai/ – /dzap/ – /w_ui/ –	192	◯	大雜匯
/dzap/ – /m_rt/ – /f_c5/ –	193	◯	雜物房
/tsu5/ – /s_rn/ – /h_rp/ – /b_ik/	194	◯	從新合璧
/g_a_/ – /g_ci/ – /t_it/ – /p_a5/	195	◯	加蓋鐵棚
/b_rt/ – /w_ak/ – /tsi_/ – /tsjy/	196	◯	筆劃次序
/s_rt/ – /m_i5/ – /j_rn/ – /s_i_/	197	◯	失明人士
/d_yt/ – /b_iu/ – /dzci/ – /m_c5/	198	◯	奪標在望
/b_ak/ – /j_uk/ – /p_ui/ –	199	◯	白玉佩
/f_uk/ – /w_ut/ – /dzit/ –	200	◯	復活節
/d_jk/ – /m_c_/ – /dzi5/ – /s_ri/	201	◯	琢磨精細
/b_ei/ – /s_ru/ – /dzak/ – /b_ei/	202	◯	備受責備
/l_in/ – /f_a_/ – /b_cu/ – /d_i5/	203	◯	蓮花寶鼎
/t_in/ – /h_ci/ – /g_u5/ – /tsi5/	204	◯	塡海工程
/m_an/ – /n_in/ – /l_ik/ –	205	◯	萬年曆

/dzyn/ – /s_jn/ – /g_an/ –	206	◯	轉瞬間
/dzrn/ – /dzat/ – /f_an/ – /k_c5/	207	◯	掙扎反抗
/h_in/ – /ngri/ – /j_y_/ – /b_rn/	208	◯	獻藝娛賓
/gwc5/ – /f_uk/ – /dzcu/ – /gwck/	209	◯	光復祖國
/tsc5/ – /h_ci/ – /s_c5/ – /t_in/	210	◯	滄海桑田
/d_ai/ – /h_j5/ – /l_ei/ –	211	◯	大鄉里
/s_j5/ – /g_a_/ – /b_it/ – /dzr5/	212	◯	商家必爭
/tsj5/ – /k_ru/ – /j_rt/ – /h_u5/	213	◯	搶購一空
/l_c5/ – /nga_/ – /p_a5/ –	214	◯	狼牙棒
/s_rm/ – /f_a_/ – /g_ci/ – /g_ap/	215	◯	深化改革
/h_rm/ – /h_ci/ – /dzu5/ – /l_j5/	216	◯	陷害忠良
/g_am/ – /l_am/ – /j_ru/ –	217	◯	橄欖油
/s_rm/ – /tsi5/ – /g_am/ – /g_ai/	218	◯	心情尷尬
/g_rm/ – /k_c5/ – /tsj5/ –	219	◯	金礦場
/d_u5/ – /n_am/ – /nga_/ –	220	◯	東南亞
/s_rp/ – /f_c5/ – /dze_/ –	221	◯	拾荒者
/dzjn/ – /s_i_/ – /dzap/ – /h_rp/	222	◯	準時集合
/h_ip/ – /l_ik/ – /t_u5/ – /s_rm/	223	◯	協力同心
/b_iu/ – /j_in/ – /h_rp/ – /g_ak/	224	◯	表現合格
/s_ip/ – /k_rp/ – /f_an/ – /w_ri/	225	◯	涉及範圍
/s_i_/ – /s_j5/ – /h_ap/ – /ngai/	226	◯	思想狹隘
/m_it/ – /f_c_/ – /h_ei/ –	227	◯	滅火器
/tsit/ – /h_rp/ – /t_ri/ – /tsci/	228	◯	切合題材
/j_it/ – /l_au/ – /f_ei/ – /s_j5/	229	◯	熱鬧非常
/d_rn/ – /s_an/ – /dzit/ – /g_i5/	230	◯	登山捷徑
/l_it/ – /k_j5/ – /dzr5/ – /b_a_/	231	◯	列強爭霸
/t_it/ – /ngru/ – /s_yn/ – /dzj5/	232	◯	鐵勾船長
/l_ck/ – /f_c_/ – /k_jy/ –	233	◯	落貨區
/dzik/ – /h_i5/ – /b_iu/ – /j_in/	234	◯	即興表現
/g_ik/ – /l_ck/ – /s_ri/ – /g_ai/	235	◯	極樂世界
/t_ck/ – /g_un/ – /l_i5/ –	236	◯	托管令
/g_ik/ – /k_ru/ – /s_ru/ –	237	◯	擊球手
/s_ik/ – /ngan/ – /s_i_/ – /g_an/	238	◯	食晏時間

Bibliography

[1] Makoto Hirayama, Eric Vatikiotis Bateson, and Mitsuo Kawato. Physiologically-based speech synthesis using neural networks. *IEICE Transactions - Fundamentals*, E76-A(11):1898–1910, 1993.

[2] J. B. Costello and F. S. Mozer. Time domain synthesis gives good quality speech at very low data rates. *Speech Technology*, 1(3):62–68, 1982.

[3] Naoto Iwahashi, Nobuyoshi Kaiki, and Yoshinori Sagisaka. Speech segment selection for concatenative synthesis based on spectral distortion minimization. *IEICE Transactions - Fundamental*, E76-A(11):1942–1948, 1993.

[4] Eric Moulines and Francis Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467, 1990.

[5] Colin Goodyear and Dongbing Wei. Articulatory copy synthesis using a nine parameters vocal tract model. In *Proceedings of IEEE 1996 International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 385–389, 1996.

[6] J. Ilse and W. Edmondson. Quasi-articulatory formant synthesis. In *Proceedings of the Third International Conference on Spoken Language Processing*, volume 3, pages 1663–1666, 1994.

[7] T. J. Sejnowski and C. R. Rosenberg. Parallel networks that learn to pronounce English text. *Complex Systems*, 1:145–168, 1987.

[8] C. R. Rosenberg. Analysis of NETalk internal structure. In *Proceedings of the Ninth Annual Cognitive Science Conference*, 1987.

[9] S. Lucas and B. Damper. Syntactic neural networks for text-phonetics translation. In *Proceedings of IEEE 1991 International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 14–17, 1991.

[10] S. M. Lucas and R. I. Damper. A connectionist approach to text-phonemics translation using syntactic neural networks. In *RNNS/IEEE Symposium on Neuroinformatics and Neurocomputers*, volume 1, 1992.

[11] Ton Weijters and J. Thold. Speech synthesis with artificial neural networks. In *IEEE International Conference on Neural Networks*, 1993.

[12] P. R. Gubbins, K. M. Curtis, and J. D. Burniston. A hybrid neural network/rule based architecture used as a text to phoneme transcriber. In *1994 International Symposium on Speech, Image Processing and Neural Networks Proceedings*, volume 1, pages 113–116, 1994.

[13] Zengjun Xiang and Guangguo Bi. A neural network model for Chinese speech synthesis. In *IEEE 1990 International Symposium on Circuits and Systems*, volume 3, 1990.

[14] M. S. Scordilis and J. N. Gowdy. Neural network control for a cascade/parallel formant synthesizer. In *Proceedings of IEEE 1991 International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 297–300, 1990.

[15] V. N. Sorokin and A. G. Miller. An articulatory-formant speech synthesizer via a neural network. In *RNNS/IEEE Symposium on Neuroinformatics and Neurocomputers*, volume 1, 1992.

[16] M. S. Scordilis and J. N. Gowdy. Speech synthesis of phonemic triplets through a neural network-controlled formant synthesizer. In *1991 International Joint Conference on Neural Networks Proceeding*, volume 2, page 1007, 1991.

[17] V. V. Kumar, S. C. Ahalt, and A. K. Krishnamurthy. Phonetic to acoustic mapping using recurrent neural networks. In *Proceedings of IEEE 1991 International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 753–756, 1991.

[18] K. M. Curtis and J. Burinston. A hybrid neural system for driving a parallel synthesizer so as to synthesize high quality accented speech. In *Proceedings of*

IEEE 1992 International Conference on Acoustics, Speech and Signal Processing, volume 2, pages 369–372, 1992.

[19] C. C. Cawley and P. D. Noakes. The use of vector quantization in neural network synthesis. In *International Joint Conference on Neural Networks*, 1993.

[20] K. M. Burniston, J. Curtis. A hybrid neural network/rule based architecture for diphone speech synthesis. In *1994 International Symposium on Speech, Image Processing and Neural Networks Proceedings*, 1994.

[21] M. G. Rahim and C. C. Goodyear. Estimation of vocal tract filter parameters using a neural network. *Speech Communication*, 9(1):49–55, 1990.

[22] M. G. Rahim, W. B. Kleijn, J. Schroeter, and C. C. Goodyear. Acoustic to articulatory parameter mapping using an assembly of neural network. In *Proceedings of IEEE 1991 International Conference on Acoustics Speech and Signal Processing*, volume I, pages 485–488, 1991.

[23] Tetsunori Kobayashi, Masayuki Yagyu, and Katsuhiko Shirai. Application of neural networks to articulatory motion estimation. In *Proceedings of IEEE 1991 International Conference on Acoustics Speech and Signal Processing*, volume I, pages 489–492, 1991.

[24] M. G. Rahim, C. C. Goodyear, W. B. Kleijn, J. Schroeter, and M. M. Sondhi. On the use of neural networks in articulatory speech synthesis. *Journal of the Acoustical Society of America*, 93(2):1109–1121, 1993.

[25] A. R. Greenwood and C. C. Goodyear. Articulatory speech synthesis using a parametric model and a polynomial mapping technique. In *1994 International Symposium on Speech, Image Processing and Neural Networks Proceedings*, volume 2, pages 595–598, 1994.

[26] T. L. Burrows and M. Niranjan. Vocal tract modeling with recurrent neural networks. In *Proceedings of IEEE 1995 International Conference on Acoustics, Speech and Signal Processing*, volume 5, pages 3315–3318, 1995.

[27] K. Shirai. Estimation and generation of articulatory motion using neural networks. *Speech Communication*, 13:45–51, 1993.

[28] M. Savic and I. H. Nam. Voice personality transformation. *Digital Signal Processing*, pages 107–110, 1991.

[29] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana. Transformation of formants for voice conversion using artificial neural networks. *Speech Communication*, pages 207–216, 1995.

[30] M. Sato, K. Joe, and T. Hirahara. APOLONN brings us to the real world: learning nonlinear dynamics and fluctuations in nature. In *International Joint Conference on Neural Networks*, 1990.

[31] Y. Ishikawa and K. Nakajima. Neural network based spectral interpolation method for speech synthesis by rule. In *The 2nd European Conference on Speech Communication and Technology Proceedings (EUROSPEECH)*, 1991.

[32] International Phonetic Association, editor. *The Principles of the International Phonetic Association.* International Phonetic Association, 1949.

[33] Secretary P. J. Roach. Report on the 1989 Kiel convention. *Journal of the International Phonetic Association*, 19(2):67–82, 1989.

[34] 葛本儀编, 《實用中國語言學詞典》, 青島出版社, 1992.

[35] 饒秉才, 歐陽覺亞 及 周無忌編, 《廣州話方言詞典》, 商務印書館, 1981.

[36] Oi kan Yeu Hashimoto, editor. *Studies in Yue dialects 1: Phonology of Cantonese.* Cambridge University Press, 1972.

[37] 李新魁, 黃家教, 施其生, 麥云 及 陳定方編, 《廣州方言研究》, 廣州人民出版社, 1995.

[38] Eric Zee. Chinese–Hong Kong Cantonese. *Journal of the International Phonetic Association*, 21(1):46–48, 1991.

[39] 周無忌 及 饒秉才編, 《廣州話標準音字彙》, 商務印書館, 1988.

[40] P. C. Ching, T. Lee, and Eric Zee. From phonology and acoustic properties to automatic recognition of Cantonese. In *1994 International Symposium on Speech, Image Processing and Neural Networks Proceedings*, volume 1, pages 127–132, 1994.

[41] 黃錫凌編, *A Chinese Syllabary Pronounced According to the Dialect of Canton*, 《粵音韻彙》, 中華書局, 1987.

[42] J. Q. Stewart. An electrical analogue of the vocal organs. *Nature*, 110:311–312, 1922.

[43] H. Dudley, R. R. Riesz, and S. A. Watkins. A synthetic speaker. *Journal of Franklin Institute*, 227:739–764, 1939.

[44] F. S. Cooper, A. M. Liberman, and J. M. Borst. The interconversion audible and visible patterns as a basis for research in the perception of speech. *Proceeding of National Academy of Science (US)*, 37:318–325, 1951.

[45] G. Fant, editor. *Acoustic Theory of Speech Production.* Mouton, 1960.

[46] Shuzo Saito and Kazuo Nakata, editors. *Fundamentals of Speech Signal Processing.* Academic Press, 1985.

[47] Howard C. Nusbaum, Alexander L. Francis, and Anne S. Henly. Measuring the naturalness of synthetic speech. *International Journal of Speech Technology*, 1:7–19, 1995.

[48] W. Wang and G. E. Peterson. Segment inventory for speech synthesis. *Journal of Acoustical Society of America*, 30:743–746, 1958.

[49] D. H. Klatt. Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, 82:737–793, 1987.

[50] N. R. Dixon and H. D. Maxwy. Terminal analog synthesis of continuous speech using the diphone method of segment assembly. *IEEE Transactions on Audio Electroacoustics*, AU-16:40–50, 1968.

[51] O. Fujimura and J. Lovins. Syllables as concatenative phonetic elements. In A. Bell and J. B. Hooper, editors, *Syllables and Segments*, pages 107–120. North-Holland, 1978.

[52] Eric Keller, editor. *Fundamental of Speech Synthesis and Speech Recognition.* John Wiley and Sons, 1994.

[53] C. P. Brownman. Rules for demisyllables synthesis using Lingua, a language interpreter. In *Proceedings of IEEE 1980 International Conference on Acoustics, Speech and Signal Processing*, pages 561–564, 1980.

[54] Yousif A. El-Imam. An unrestricted vocabulary arabic speech synthesis system. *IEEE Transactions on Acoustics Speech and Signal Processing*, 37(12):1829–1845, 1989.

[55] P. Bhaskararao, S. J. Eady, and J. H. Esling. Use of triphones in demisyllables based speech synthesis. In *Proceedings of IEEE 1991 International Conference on Acoustics Speech and Signal Processing*, volume II, page S7.28, 1991.

[56] L. R. Rabiner, R. W. Schager, and J. L. Flanagan. Computer synthesis of speech by concatenation of formant-coded words. *Bell System Technical Journal*, 50:1541–1558, 1971.

[57] Joseph P. Olive and Lloyd H. Nakatani. Rule-synthesis of speech by word concatenation : a first step. *Journal of the Acoustical Society of America*, 55(3):660–666, 1974.

[58] T. Dutoit and H. Leich. MBR-PSOLA : Text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication*, 13:435–440, 1993.

[59] H. Valbret, E. Moulines, and J. P. Tubach. Voice transformation using PSOLA technique. *Speech Communication*, 11:175–187, 1992.

[60] Tomohisa Hirokawa, Kenzo Itoh, and Hirokazu Sato. High quality speech synthesis system based on waveform concatenation of phoneme segment. *IEICE Transactions - Fundamental*, E76-A(11):1964–1970, 1993.

[61] H. Dudley. The vocoder. *Bell Laboratories Record*, 17:122–126, 1939.

[62] J. L. Flanagan and R. M. Golden. Phase vocoder. *Bell System Technical Journal*, 45:1439–1509, 1966.

[63] J. N. Holmes. The JSRU channel vocoder. *Proceedings of IEEE*, 127:53–60, 1980.

[64] Walter Lawrence. The synthesis of speech from signals which have a low information rate. In W. Jackson, editor, *Communication Theory*, pages 460–469. Butterworths, 1953.

[65] D. H. Klatt. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67:971–995, 1980.

[66] J. Kelly and L. Gerstman. An artificial talker driven from phonetic input. *Journal of the Acoustical Society of America Supplement 1*, 33:s35, 1961.

[67] J. N. Holmes, I. G. Mattingly, and J. N. Shearme. Speech synthesis by rule. *Language Speech*, 7:127–143, 1964.

[68] I. G. Mattingly. Synthesis by rule of prosodic features. *Language Speech*, 9:1–13, 1966.

[69] F. Itakura and S. Saito. Analysis-synthesis telephony based on the maximum likelihood method. In *Proceedings of the 6th International Congress Acoustics, Tokyo*, pages C–5–5, 1968.

[70] B. S. Atal and S. L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50:637–655, 1971.

[71] J. Makhoul. Spectral analysis of speech by linear prediction. *IEEE Transactions on Audio Electroacoustics*, AU-21:140–148, 1973.

[72] John D. Markel, editor. *Linear Prediction of Speech.* Springer-Verlag, 1976.

[73] R. Wiggins. An integrated circuit for speech synthesis. In *Proceedings of IEEE 1980 International Conference on Acoustics, Speech and Signal Processing*, pages 398–401, 1980.

[74] Nelson Morgan, editor. *Talking Chips.* McGraw-Hill, 1984.

[75] H. K. Dunn. The calculation of vowel resonances, and electrical vocal tract. *Journal of the Acoustical Society of America*, 22:740–753, 1950.

[76] K. N. Stevens, S. Kasowski, and G. Fant. An electrical analog of the vocal tract. *Journal of the Acoustical Society of America*, 25:734–742, 1953.

[77] K. N. Stevens and A. S. House. Development of a quantitative description of vowel articulation. *Journal of the Acoustical Society of America*, 27:484–493, 1955.

[78] J. L. Flanagan, editor. *Speech Analysis Synthesis and Perception.* Springer, 1972.

[79] C. H. Coker. Speech synthesis with a parametric articulatory model. In *1968 Speech Symposium, Kyoto*, pages A–4, 1968.

[80] R. Teranishsi and N. Umeda. Use of pronunciation dictionary in speech synthesis experiments. In *Reports of the Sixth International Congress on Acoustics, Tokyo*, volume 2, pages 155–158, 1968.

[81] E. Matsui, T. Suzuki, N. Umeda, and H. Omura. Synthesis of fairy tales using an analog vocal tract. In *Reports of the Sixth International Congress on Acoustics, Tokyo*, volume B, pages 159–162, 1968.

[82] J. L. Kelly and C. C. Lochbaum. Speech synthesis. In *Proceedings of the Fourth the Sixth International Congress on Acoustics*, volume G42, pages 1–4, 1962.

[83] C. H. Coker, N. Umeda, and C. P. Brownman. Automatic synthesis from ordinary English text. *IEEE Transactions on Audio Electroacoustics*, AU-21:293–297, 1973.

[84] N. Umeda. Linguistic rules for text to speech synthesis. *Proceedings IEEE*, 64:443–451, 1976.

[85] J. L. Flanagan, K. Ishizaka, and K. L. Shipley. Synthesis of speech from dynamic model of the vocal cords and vocal tract. *Bell System Technical Journal*, 54:485–506, 1975.

[86] J. L. Flanagan and L. Landgraf. Self-oscillating source for vocal tract synthesizer. *IEEE Transactions on Audio Electroacoustics*, AU-16:57–64, 1968.

[87] Ishizaka K. and J. L. Flanagan. Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell System Technical Journal*, 51:1233–1268, 1972.

[88] Paul Mermelstein. Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53:1070–1082, 1973.

[89] C. H. Coker. A model of articulatory dynamics and control. *Proceedings IEEE*, 64:452–459, 1976.

[90] M. H. L. Hecker. Studies of nasal consonants with an articulatory synthesizer. *Journal of the Acoustical Society of America*, 34:179–188, 1962.

[91] P. Meyer, R. Wilhelms, and H. W. A. Strube. Quasiarticulatory speech synthesizer for german language running in real time. *Journal of the Acoustical Society of America*, 86:523–539, 1989.

[92] Man Mohan Sondhi and Juergen Schroeter. A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Transactions on Acoustics Speech and Signal Processing*, 35:955–967, 1987.

[93] B. S. Atal. Determination of the vocal tract shape directly from the speech wave. *Journal of the Acoustical Society of America*, 47:65, 1970.

[94] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *Journal of the Acoustical Society of America*, 63:1535–1555, 1978.

[95] Juergen Schroeter and Man Mohan Sondhi. Techniques for estimating vocal-tract shapes from the speech signals. *IEEE Transactions on Acoustics Speech and Signal Processing*, 1(2):133–150, 1994.

[96] Tomaso Poggio and Federico Girosi. Network for approximation and learning. *Proceedings of IEEE*, 78:1481–1497, 1990.

[97] G. Widrow and M. E. Hoff. Adaptive switching circuit. *IRE Western Electronic Show and Convention : Convention Record*, pages 96–104, 1960.

[98] W. K. Lo and P. C. Ching. Phone-based speech synthesis with neural network and articulatory control. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, pages A–929, 1996.

[99] G. Cybenko. Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, 1989.

[100] K. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2:183–192, 1989.

[101] B. Moore and T. Poggio. Representations properties of multilayer feedforward networks. In *Abstracts of the First Annual INNS Meeting*, page 502. Pergamon Press, 1988.

[102] S. Y. Kung, editor. *Digital Neural Networks.* Prentice-Hall, 1993.

[103] Jr. John R. Deller, John G. Proakis, and John H. L. Hansen, editors. *Discrete-Time Processing of Speech Signals.* Macmillan Publishing, 1993.

[104] S. Chen, C. F. N. Cowan, and P. M. Grant. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, 2(2):302–309, 1991.