Domain-Optimized

Chinese Speech Generation

馮恬瑩

FUNG Tien Ying

A Thesis

Submitted in Partial Fulfilment of the Requirements for the Degree of

Master of Philosophy

in

Systems Engineering and Engineering Management

August 2001

Domain-Optimized Chinese Speech Generation

by
馮恬瑩
FUNG Tien Ying

A Thesis
Submitted in Partial Fulfillment of the Requirements for the Degree of
Master of Philosophy
in
Department of Systems Engineering and Engineering Management

Abstract

The goal of this thesis is to optimize the naturalness and intelligibility of corpus-based syllable concatenation for limited-domain synthesis in Chinese. We first test the feasibility of the approach by developing a speech generation framework in a constrained domain. The framework is applicable across two Chinese dialects, Mandarin and Cantonese. To obtain highly natural speech output in our corpus-based syllable concatenation approach, coarticulatory context is taken into consideration by the use of distinctive features. The framework is also demonstrated to be scalable and portable with several enhancements made in the architecture of the framework and the creation of large scale domain-optimized wavebank. The next step of our work is the investigation on tonal contexts, which is also crucial to the naturalness of a syllable-based concatenated speech. Our investigation involves two major component: (i) to find out the relative importance of left and right tonal contexts towards naturalness; (ii) to develop a backoff unit selection scheme for tonal context in case of missing tonal variants. Such investigation helps to improve the quality of speech generated in more complex domains.

Thesis Supervisors: Helen M. Meng, Boon-Toh Low
Title: Associate Professor, Assistant Professor

摘要

本論文旨在探討在中文的受限範疇中優化語音合成自然度及可理解性的方法. 我們首先在被約束的範疇內建立一個語音合成結構,以測試我們所提議的方法的可行性. 這是一個可被採用於兩種不同中文方言的結構,包括普通話及廣東話. 爲了提高語音輸出的自然度,我們在這文本音節拼接方法中利用了分節音前後關係的不同特點. 透過語音合成結構的改進及大規模範疇優化語音庫的建立,我們顯示了這結構在不同範疇中的可變比例性及可移植性. 然後,我們探討另一個影響到語音音節拼接的自然度的重要因素–音調的前後關係. 其中包括: 一. 找出左右音調分別對自然度的相對重要性, 二. 建立一個音節單位選取方案, 當語音庫內無所需的音調單位時, 我們可以選出最能取代的單位. 這探討改進了在複雜範疇內合成語音的質素.

Acknowledgments

The two year of research has been a long road to me. Throughout these two years, I have encountered ups and downs, happiness and unhappiness. Unhappy are the times when there are obstacles in my research that stopped me from moving forward, that through try and retry, I still could not find the right solution. Happy are the times when I am sharing my everyday research life with my peer classmates and the research staffs in the Human-Computer Communications Laboratory.

I would like to thank my two supervisors, Professor Helen Meng and Professor Boon-Toh Low for their support. Thanks to Professor Helen Meng, I learnt the meaning of persistence. She shared with me the joy when I have improvements in my research, and in my hours of darkness, she gave me encouragement and research guidance. Although there are times that I failed to meet her expectations, she never gave up on me. On the contrary, she showed that she has confidence in me, which supported me to go on. She not only helped me with my research, but also shared with me her life experience. Under her supervision, I learnt more than I expected.

I thank Professor Boon-Toh Low for the degree of freedom he gave me, and also his suggestions for my research that helped me to see broader and deeper into the problems. I am also grateful for his patience and encouragement towards me.

I would also like to express my gratitude to my thesis committee, Professor Meng, Professor Low, Professor Chris Yang, Professor Jeffrey Yu, Professor Lin-Shan Lee from National Taiwan University and Dr. Chilin Shih from Bell Laboratory, for their time and interest.

As I said, this two year is not easy for me. I would not have made it without the support from my friends. I would like to say thanks to Carmen, Sally, Connie and Yuk, for they are my closest comrades, we worked through the hardship together. I would also like to thank the HCCL fellows: Ada Luk, Timmy, Kin, KFan, Tony, Ida, Brenda, Julia, Chat, Michael Lo, Sunny, Tall-guy Edmond and Silvia for their help, support, encouragement and laughters provided to me within the two years, and the sweet and wonderful friendship they gave me. I also want to thank my past FYP teammates: Angel, Brenda, Jessica and FYP teams of Spgen99: Lui Min, Cherie, Ah Ling, for their help in creating the wavebanks in the FOREX and ISIS domains. Others I would like to thank include Coral, Philip, Patrick, Bunny, ah Su, Wai-Ip, Kuen-

Chung, Wai-Kit Lo, Yiu-Wing Wong, Phyllis Leung, Mandy Tsoi, Iris Leung and Aggie Yip.

Finally, I wish to express my deepest gratitude to my family: Mom, Dad and my brother Walter, as well as my special one, Yuk, for the love and care they devote to me. "I can take all the madness the world has to give, but I won't last a day without you.". They are the wind beneath my wing. I love them very much.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Conversational interfaces is a breakthrough for human-computer interaction. Conversational interfaces is becoming increasingly desirable as it offer ease-of-use of human-computer communication: speech input is easer than keyboard input by typing; speech output augments the visual display on the computer screen. Conversational interfaces further extend the limit of the usability of computers by supplementing the in-sufficiency of the traditional interaction style between human and computer.

Speech generation technology is an essential component in conversational interfaces. It offers an alternate channel for information delivery in a hands-busy, eyes-busy environment. It can also be plugged into dialog systems for response generation, e.g for systems like "CU FOREX" [3] and "ISIS" [4, 5]. Potential applications include aids for the handicapped, mobile computing, speech-to-speech translation, email reader and news reader.

1.1 General Trends on Speech Generation

The ultimate goal of speech generation is to achieve complete *flexibility* at perfect *speech quality*. Flexibility refers to the ability to generate speech for any given text. Speech quality refers to naturalness, which is how well the synthesized speech sounds like real human speech; and intelligibility, which is how easily the synthesized speech can be understood. However, there is a tradeoff between flexibility and speech quality in the development of speech generation technology. Figure 1.1 shows the development tradeoff schematic.

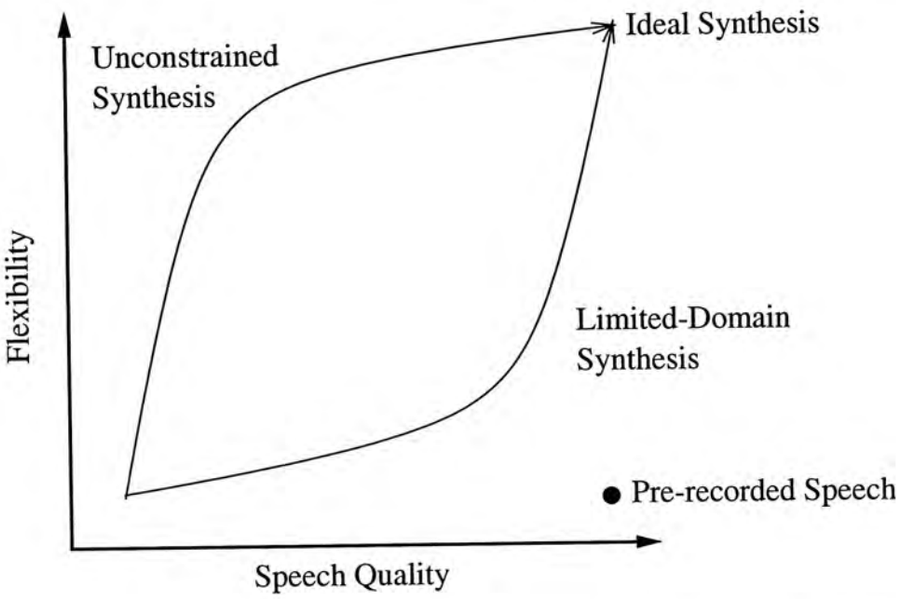


Figure 1.1: Speech generation development tradeoff schematic. (This figure is borrowed from [1])

Two major approaches, namely *unconstrained synthesis* and *limited-domain synthesis* are developed in different directions based on the priority set for the flexibility and speech quality. The upper line describes the development

trend for unconstrained synthesis, in which flexibility is achieved first and then speech quality is improved. The lower line describes the development trend for limited-domain synthesis, in which highly natural and intelligible speech quality is achieved first and then flexibility is improved. The black dot at the bottom shows an extreme case of limited-domain synthesis using pre-recording speech. Unconstrained synthesis has high flexibility but it is difficult to be able to sustain a high quality for speech output under all conditions. Synthesis in restricted domains provides a scope within which the speech quality can be optimized, but then by definition we are compromising on flexibility. However, by combining more and more domains, limited-domain synthesis can achieve the effect of domain independence.

1.2 Domain-Optimized Speech Generation in Chinese

Domain optimization in speech generation refers to developing corpus tailor to a given domain, so that the naturalness and intelligibilty can be optimized. This thesis aims to develop a framework for domain-optimized generation across Chinese dialects. We focus on Mandarin and Cantonese. Mandarin is the most widely spoken Chinese dialect and it is used by over 720 million people [6], and Cantonese is a major dialect used by over 60 million people mainly in Southern China [7]. They are also the two most common dialects used in our region.

We adopt a corpus-based concatenative approach in our framework. Our approach takes advantage of the monosyllabic nature of Chinese and the concatenation is based on the syllable units. To optimized the naturalness of output speech, coarticulatory and tonal context are taken into consideration. Our framework captures coarticulation by means of distinctive features. Prosodic effects are indirectly handled during recording corpus development. However, this kind of prosodic modeling is only feasible in domains which are relatively constrained, such as the foreign exchange domain. The framework is enhanced so that it is scalable and portable to more complex domains.

In order to obtain highly natural speech quality in more complex domains, influence of tonal context needed to be handled explicitly. Hence we conduct a study on the influence of tonal context. Our investigation focus on the relative importance between left and right tonal contexts. Thereafter we developed a unit selection scheme to minimize the tonal effects in generated speech.

1.3 Thesis Organization

The rest of this thesis is organized as below:

In Chapter 2, we introduce the linguistic properties of Chinese, as well as various approaches and previous work done in the area of speech generation.

Chapter 3 presents a feasibility test of our corpus-based syllable concatenative approach by applying our speech generation framework in the

foreign exchange domain. We describe the optimization of speech quality by capturing coarticulation using distinctive features, and also the creation of our domain-optimized wavebank (i.e. the library of syllable units for concatenation). We also describe the methodology of syllable unit selection for concatenation, followed by a listening to demonstrate highly natural speech output produced by the framework.

In chapter 4, we describe the scalability and portability of our framework. We migrate the framework to the stocks domain by plugging it into a spoken dialog system, ISIS [5]. We describe the enhancements made on the framework, which include the use of XML, a alternate approach for large-scale wavebank creation, and the application of the energy normalization technique to improve output speech quality.

We move on to describe our study in the influence of tonal context in chapter 5. We present our investigation in the relative importance of left and right tonal contexts, as well as a backoff scheme for unit selection in terms of tonal context. The backoff scheme suggests a ranked list of substitutes for missing syllable units. We will show the validity of the unit selection scheme by a listening test.

Finally, chapter 6 gives a summary of this thesis, as well as the contribution made and the future directions of our research in speech generation technology.

Chapter 2

Background

Our approach for speech generation is corpus-based syllable concatenation in Chinese. In this chapter, we will first introduce some Chinese linguistic and phonological properties. Afterwards, we will give an overview on previous developments in speech generation. It involves three main approaches, namely articulatory synthesis, formant synthesis and concatenative synthesis, as well as some previous work on this research area. Lastly, we will describe our approach in Chinese speech generation.

2.1 Linguistic and Phonological Properties of Chinese

Each Chinese character is pronounced as a syllable with a tone. The syllable itself is referred to as a *base syllable*; and when it comes with a tone, it is

called a *tonal syllable.* The pronunciations of Chinese characters can be fully represented with tonal syllables. In Mandarin, there are around 1400 tonal syllables and reduced to 418 if tone is not taken into consideration [8]. In Cantonese, there are around 1700 tonal syllables and 619 base syllables [9].

A base syllable can be divided into a optional syllable initial and a syllable final. The initial is a (optional) consonant onset and the final consists of a vowel nucleus and an optional consonant coda [7]. Onsets and codas cause different *articulations* at syllable boundaries, which is an important consideration in syllable-based concatenative synthesis in Chinese. Another important consideration is *tone.* The realization of a syllable's tone can be affect by the tones of its neighboring syllables. The same syllable can carry entirely different linguistic meanings when it is produced in different tones.

2.1.1 Articulation

As mentioned, onsets and codas represents the sounds of a syllable at its boundaries. The differences in sounds are caused by different movements of articulator , such as lips, tongue and jaw. The movements are referred to as *articulation.* In Chinese syllable-based concatenative synthesis, the consideration of articulatory features accounts greatly for the naturalness and intelligibility of synthesized speech.

In our approach, we use distinctive features to model coarticulatory characteristics. Distinctive feature is a set of linguistic units to distinguish a set of maximally close phonemes. For example, syllable onsets with the initials

/b, p, m, f, w/ are labelled with the feature LABIAL; syllable codas with finals ending in /g, k/ are labelled VELAR.

2.1.2 Tones

Mandarin has four basic tones, they are high level, high rising, dipping/falling and high falling. There is also a additional tone, namely the neutral tone. It is the phenomenon that a syllable is reduced and carries on lexical tones. In this thesis, the investigation of tone focus only on Cantonese tones, therefore we will not introduce Mandarin tone system in detail.

Cantonese has a more complex tone system. There are nine tones, upper level, upper rising, upper going, lower level, lower rising, lower going, upper entering, middle entering and lower entering (see Figure 2.1).

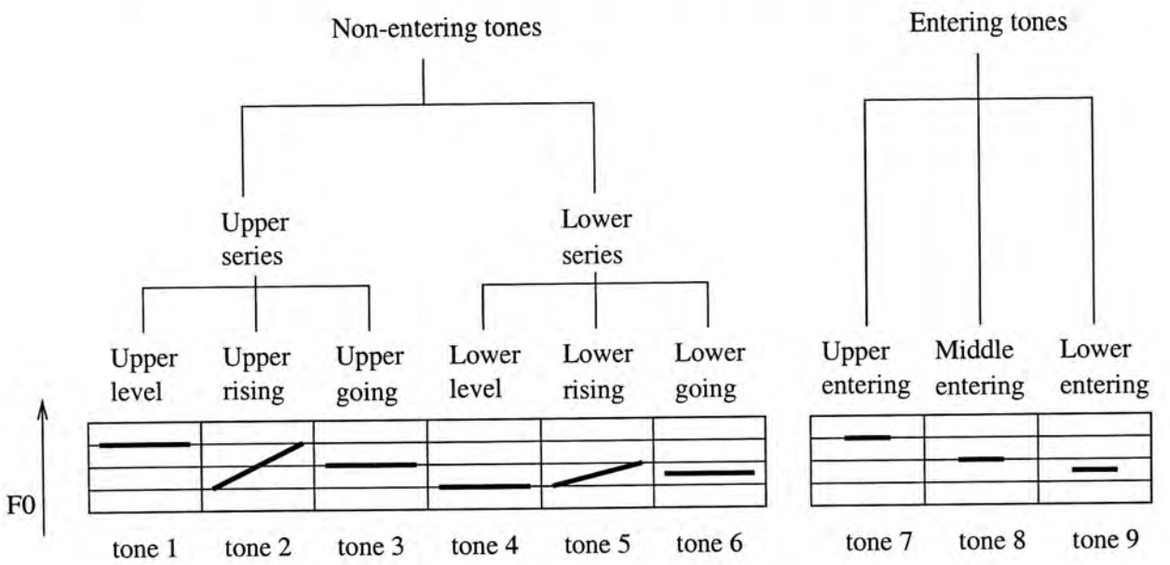


Figure 2.1: The Cantonese nine-tone system.

Six out of nine tones are categorized as non-entering tones, the other three of them are categorized as entering tones. The difference between the categories of entering and non-entering tone is that entering tones are shorter in duration, all of those syllables with entering tones always end with a stop consonant /p/, /t/ or /k/. As the three entering tones have identical tone shape and height as some of the non-entering tones, the nine-tone system can be simplified into a six-tone system, in which tone 1 merges with tone 7, tone 3 merges with tone 8 and tone 6 merges with tone 9. In this thesis, we work with the six-tone system for Cantonese as we only focus on the shape and height in our tone investigation.

2.2 Previous Development in Speech Generation

Speech generation has a long history since the first connected-speech synthesizer, "Voder" was developed in 1939 by Dudley [10]. Along the development history, there has been three major synthesis approaches, namely *articulatory synthesis*, *formant synthesis* and *concatenative synthesis*. The former two synthesis approaches were developed based on the source-filter theory of speech production proposed by Fant in the 1960s [11]. The source-filter model suggests that speech synthesis can be divided into two parts, one is the model of energy source and the other is the model of the vocal tract transfer function. As the modeling depends heavily on manually derived

rules, the two synthesis approaches are also classified as rule-based synthesis. Concatenation synthesis, in contrast to articulatory synthesis and formant synthesis, is classified as data-driven synthesis. In this approach, speech is generated by concatenating real speech segments. It is advantageous to use this approach in limited-domain synthesis, as it can give very high quality speech with a small number of recorded segments in a given domain. In the following sections, we will describe the three synthesis approaches in detail.

2.2.1 Articulatory Synthesis

Articulatory synthesis attempts to produce speech by constructing mathematical models of the articulator movements in the vocal system. The schematic diagram of vocal system is shown in Figure 2.2.

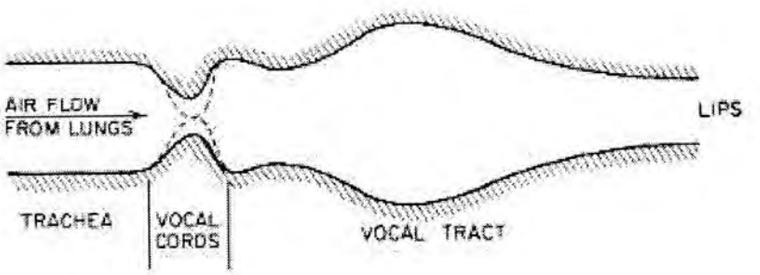


Figure 2.2: Schematic representation of the vocal system (this figure is borrowed from [2]).

The mathematical models are developed base on some physical principles and simplifying assumptions about the physical properties of the vocal system. As the knowledge of articulatory phonetics is insufficient, most work on

articulatory synthesis are only based on simplified two-dimensional models. Three-dimensional models are developed only recently [12]. Although it was once believed that this may be the best approach to the synthesis of natural sounding synthetic speech [13], such approach is not commonly applied applied in real-time speech generation systems [14]. The challenge in the adoption of articulatory synthesis is that it is extremely difficult to formulate the articulator motion and to compute the acoustic properties from the shape of vocal tract.

2.2.2 Formant Synthesis

Formant synthesis bypasses the computational complexity in modeling articulator movements and acoustic properties, by modeling them from their speech spectral representation. The spectral representation refers to the trajectories of formants in a spectrogram [15]. Formants are the major resonant frequencies of the vocal tract. These frequencies change as the configuration of the vocal tract changes.

The approach of formant synthesis produce speech by modeling the changes in the frequency and amplitude in the formants for the vowels, including the shape of the transitions of those formants in order to synthesize the consonants. Formant synthesis can produce high quality speech output if the trajectories of formants are modeled precisely. However, hundreds of rules that correctly describe the trajectories are required [16]. As the task complexity is too high, work has been done on formant synthesis by trying to

reduce the number of control parameters needed [17]. Formant synthesis is applied in many commercial speech synthesizers because of its compact size [18]. However, the task complexity is still a challenge for producing perfect quality speech.

2.2.3 Concatenative Synthesis

The complexity in concatenative synthesis is lower when compared to the two previous approaches. Concatenative synthesis involves taking real recorded speech, cutting it into segments, and concatenating these segments back together during synthesis. As it generates speech by concatenating human produced speech segments, this approach can produce speech with high naturalness and intelligibility.

The challenge in this approach is the discontinuity at concatenation points because of coarticulation and prosody mismatch [19]. If the discontinuities are prominent, concatenation synthesis can result in poor quality speech despite the naturalness within each concatenation unit.

There are four crucial issues in concatenative synthesis: the choice of concatenation unit, recording corpus development, unit selection for concatenation and prosody modification.

1. **Choice of Concatenation Units**

   Concatenation units refers to the recorded segments for concatenation. It can be a phone, a diphone, a demisyllable, a syllable, a phrase or

variable-length units.

Phone-based concatenative synthesis has been difficult because of the strong contextual variation in the acoustic realization of each phone. There are problems in storing exhaustive variants of each phone (i.e. allophones) and selecting the appropriate units for concatenation. Phone-based concatenation provide high flexibility because of the small size of phone. Nevertheless, since the unit is small, number of concatenations is inevitable large, that leads to unnatural resulting speech. Previous work in phone-based Chinese speech generation can be found in [20].

Diphones are phone to phone transitions that begin in the middle of the stable state of a phone and end in the middle of the following one [21]. Early work on diphone is reported in 1977 by Olive [22]. Successful work using diphone-based concatenative synthesis includes the MBROLA project developed by the TCTS laboratory [21], and the female-voice Mandarin synthesizer developed by Bell Laboratories [23]. The advantage of diphone-based concatenation is that diphones involve most of the transitions and coarticulations between phones, and there is a relatively small set of diphones. However, since large distortions in formant trajectories may occur in two units obtained from different context, there is no guarantee for perfect matches in the concatenation points.

Work has also been done with demisyllable as concatenation units. Each Chinese syllable can be divided into two demisyllables of initial and final. Due to the monosyllabic nature of Chinese, demisyllable-based concatenation is more preferable than phone-based concatenation in Chinese synthesis [24, 25].

As computer speed and storage increase at Moore's pace, researchers tends to favour larger and less context-sensitive units, such as syllables and phrases [26], as they give higher quality speech. It is advantageous to perform syllable-based concatenative synthesis in Chinese speech generation due to the mono-syllabic nature of Chinese. Past work can be found in [8, 27, 28, 29, 30, 31, 32, 33, 34, 35] Syllable-based concatenative synthesis is also desirable as the coarticulations across syllable boundaries are weaker than that between phones within a syllable [36].

Phrase level concatenation is capable of producing even higher quality speech because of the reduced points of concatenation. However, it is highly inflexible. Therefore, phrase units are often used together with syllable units, where phrase units are used for frequently occurring words and syllable units are used as basic concatenation units [37]. In order to maximized both naturalness and flexibility, the use of variable-length units became popular in concatenative synthesis [1, 38, 39].

Taking the advantage of the mono-syllabic natural of Chinese, syllable and phrase level concatenation is adopted as our approach in Chinese speech synthesis, due to its ability to achieve high naturalness with reasonable flexibility.

2. **Recording Corpus Development**

   Since concatenative synthesis aims to generate speech via concatenation of real acoustic segments, it is essential to have the entire vocabulary recorded. There are two common approaches on recording corpus development. One is to carefully design a set of utterances to cover units in various context [40]. The other is using large corpus of single speaker speech data with orthographic transcription [41, 42].

   For limited-domain synthesis, especially for response generation for information delivery, the first approach is preferable. It is desirable to record corpus in carrier phrase of which identical to the response need to be generated [43], so that the coarticulatory and prosody environment of the units can be preserved best. Therefore, this approach is applied on our response generation framework. We developed an algorithm to ensure full coverage of acoustic units in the scope of domain with a maximally compact set of recording prompts.

3. **Unit Selection for Concatenation**

   Unit selection is a process which ensures that the best unit sequence is chosen to give a optimal quality speech. An approach for unit selection

scheme is by numeric measurements for a concatenation of acoustic units so as to minimized distortion at the junctions. It is achieved with a distance metric which consist of two costs, namely the unit cost and the transition cost [1, 44, 45, 46, 47, 48]. Unit cost is a combination of the coarticulation cost and the prosody cost, which are the measures for coarticulation and prosody. Transition cost incorporates the continuity measures for coarticulatory and prosody [49]. The approach is designed for large generic speech corpus.

In this thesis, the wavebank of acoustic units is small in size as limited-domain synthesis is performed. Therefore, a much simpler algorithm of distinctive feature matching is applied for unit selection to meet the same purpose of ensuring optimal quality speech.

4. **Prosody Modeling**

Prosody relates to pitch level, energy and duration. Speech output by concatenative synthesis may have poor quality because of problematic prosodic realizations. A widely used approach for prosody modeling is prosody modification. That is to adjust the pitch contour, energy and duration of the generated speech as a post-process. This process helps to correct the prosody of the speech output to match the target prosody. A commonly used prosody modifications techniques is Pitch Synchronous Overlap and Add (PSOLA) [50, 51]. A variation from PSOLA is Time-Domain Pitch Synchronous Overlap and Add

(TD-PSOLA) [52, 53] which solved the problem of phase mismatches introduced in PSOLA, however, perceived noise may occur. Prosody modification can also done by using ToBI labelling [54]. ToBI labelling use abstract representation to mark the prosodic target, such as "H" for high pitch level, "L" for low pitch level. It includes also diacritics to indicate various intonation functions.

Prosody modification are also included in Chinese speech synthesizers. However, since a syllable's tone can potentially be distorted by its neighboring tones, the result from prosody modification will still sound erroneous if the target prosody is determined with the pitch contour of lexical tones. The consideration of tone variation is important. Attempts to capture tonal variations include the use of statistical model and tone coarticulation rules. Past investigation in rule based tone variation modeling can be found in [8, 55, 56]. Past work by Bell Laboratories on tone modeling involves the use of Stem-ML tags [57].

Most of the past work obtains a better prosody realization of generated speech after prosody modification. However, this may also degrade the quality of human produced speech segments due to signal processing. In our approach, we avoid the altering the pitch contour of the output speech. We capture syllable units with various tone variations directly from the recording corpus and a unit selection scheme is developed to select syllable units with the best matching tone shape. As past work on Cantonese tones investigation are relatively sparse,

our work focus on Cantonese tones.

2.2.4 Existing Systems

There has been remarkable developments in this field by various research groups. Some of the leading work in speech generation is listed below:

- The Bell Laboratories Text-to-Speech System

  Bell Laboratories of Lucent Technologies has very a long history with the research of speech synthesis, since the demonstration of VODER in 1939. Early work on articulatory synthesis was done in the mid-to-late 1960s [18]. The development of concatenation synthesis was started by Olive in the mid 70s [22]. The approach has been adopted for all the text-to-speech system development since then. Most of their work focus on American English generation, and more recently, they expending the language set to other languages including Mandarin Chinese.

  The majority of concatenation units are diphones. The units are chosen based on various criteria that include spectral discrepancy and energy measure. Context-sensitive allophonic units or triphones are used for the consideration of contextual or coarticulory effects. Algorithms for automated optimal element selection and cut point determination are developed to minimize spectral discrepancies between elements and maximize the coverage of required elements. They also developed unit selection and concatenation modules to retrieve the necessary units,

assign new durations, pitch contours and amplitude profiles [58]. Bell Laboratories also have works on Chinese tone modeling, a markup language called "Stem-ML" is developed by Shih [57] to model tonal variations.

- CSTR Speech Synthesis System

  Festival [59] is a general multi-lingual speech synthesis system developed at CSTR at the University of Edinburgh by Black and Taylor, in co-operation with CHATR, Japan. The system currently support English (British and American), Spanish and Welsh text-to-speech synthesis. It is diphone based with techniques of residual excited LPC and PSOLA. Besides Festival, CSTR also developed Concept-to-Speech synthesis system through the SOLE project in the domain of museum guide (ILEX), in which phonological trees are applied for unit selection [60].

- The MIT Speech Synthesis System

  Recent work on Chinese speech synthesis system in the Spoken Language System Group at MIT includes TD-PSOLA concatenative synthesis using diphones by Yi [61] and variable-length units concatenation by Yi [1]. In TD-PSOLA diphone based concatenation, pitchmarks and phone boundaries are used to excise diphones from a given utterance. A search algorithm is applied to match diphones to the specified phone sequence. The variable-length units concatenation is

a concept-to-speech response synthesis achieved by word- and phrase-level concatenation. The units are carefully prepared in the precise prosodic environment using carrier phrases as vehicle. The unit corpus is search during synthesis using a Viterbi search with the use of unit cost function and transition cost function. The concept-to-speech framework has proven to be able to generate natural speech and is more preferable than diphone synthesis.

- Mandarin Synthesis System development at National Taiwan University

  National Taiwan University has research on Chinese speech generation since 1980s. They have developed a Mandarin text-to-speech system using syllable concatenation approach [8]. The text-to-speech synthesis is performed based on a set of synthesis rules, which are derived from the acoustic properties of Mandarin. The synthesis rules focus on prosody modeling, they include tone concatenation rules, tone sandhi rules, stress rules, intonation patterns, syllable duration rules, pause insertion rules, and energy modification rules. These rules are proven to be very useful in improving naturalness and intelligibility of the output speech quality. They also help in understand the tone variation behavior and the characteristics of continuous Mandarin speech.

- Mandarin Synthesis System development at Microsoft Research China

  Microsoft Research China has been developed a Mandarin speech syn-

thesizer with the approach of non-uniform concatenative synthesis with tonal syllables as the basic concatenation units by Chu [37]. The concatenation units are extracted from a large speech corpus that covers all acoustic units and most of their variations. The instance of units are classified into categories of prosody features. A multi-tier unit selection algorithm is used, where the first tier is to select leaf nodes on the indexing tree of speech corpus, the second is to prune the initial lattice of unit by a contextual distance weight sum, the last is to find the best path by minimizing concatenation cost. The work resulted in very natural and fluent synthesized speech.

- The Cantonese Speech Synthesizer in CUHK

  CUHK Speech and Language Technology Group has been working on speech synthesis for Cantonese, which is a major Chinese dialect used in our region. The synthesis approaches used includes phone-based concatenative synthesis by Lo et al. [62] and syllable-based concatenative synthesis by Lee et al. [7]. In the phone-based approach, neural network is employed for the generalization of the phone templates during synthesis. It has articulatory control provided from simplified articulatory space input parameters. The network approach is chosen for its non-linear mapping of the relationship between the articulatory space parameters and the spectral information of the speech signal. Result from an informal listening test shows that the approach gives a fair speech quality.

In the syllable-based approach, the technique of TD-PSOLA is used for prosodic modification. The technique was first adopted in Cantonese synthesizer developed at CUHK by Chu and Ching in 1997 [53, 63]. The work is a pioneer study on prosody control for Cantonese speech synthesizer. The prosodic control in the work refers to the control of segment-level temporal structure and variation of fundamental frequency. Based on statistical derivation from a large speech database, a set of prosodic rules is established to improve the naturalness of output speech.

2.3 Our Speech Generation Approach

Our approach for Chinese speech generation is corpus-based syllable concatenation. We chose the syllable as our basic concatenation unit because of the mono-syllabic nature of Chinese. We developed a generate-and-filter algorithm to create recording corpus within the scope of domain. It aims to cover all coarticulatory variants of syllable units in a compact set of automatically generated recording prompts. We model coarticulatory context by the use of distinctive features. We concatenate syllable units sequentially from left to right, and our unit selection approach simply enforces units with matching left and right coarticulatory context is chosen. For prosodic modeling, we investigate the influence of left and right tonal contexts, and developed a backoff unit selection scheme to select syllable units with best matching tonal

context when a perfect match is absent. In this way, no signal processing is needed to alter the pitch contour of the generated speech.

Chapter 3

Corpus-based Syllable Concatenation: A Feasibility Test

In this chapter, we will describe our approach designed for Chinese speech generation, which is based on a syllable concatenation technique. Our approach can optimize the speech quality of the generation output within the scope of a given domain. We aim to demonstrate the applicability of this approach for the foreign exchange (FOREX) domain, as well as for two key Chinese dialects – Cantonese and Mandarin.

We adopt tonal syllables as our basic unit for concatenation. This is because Chinese is by nature monosyllabic. A finite set of tonal syllables provide complete phonological coverage of the Chinese language. Chinese is

also a tonal language, i.e. each syllable carries a tone, and a given syllable with different tones may convey different lexical meaning. Mandarin has 1400 tonal syllables in its phonological space, while Cantonese has 1700. Syllable acoustics in continuous speech are affected by their neighboring syllables – an effect known as *coarticulation.* In our approach, we model coarticulation by means of *distinctive features.* Coarticulation modeling is a critical factor in achieving a high degree of intelligibility and naturalness in the generated speech.

We have selected the FOREX domain for feasibility demonstration. The domain is well-suited for Hong Kong as our region is one of the major foreign exchange trading centers in the world. The source of our financial data is the real-time Reuters satellite feed. The foreign exchange data are very dynamic and we need to be able to verbalize the information as well as generate speech output instantaneously upon demand. This calls for real-time speech generation technology. Another desirable characteristic of the FOREX domain is its simplicity. There are only several major domain-specific information categories: date, time, currency names, and exchange rates. The domain is too complex for speech generation by pre-recorded phrases due to combinatorial explosion. However, the domain should be simple enough with constrained variations in prosodics and coarticulatory contexts. Hence FOREX is very desirable as an application domain for feasibility demonstration of our syllable-based concatenative approach for Chinese speech generation.

3.1 Capturing Syllable Coarticulation with Distinctive Features

There is a certain structure for Chinese syllables. Each Chinese syllable consists of an onset (optional), a nucleus and a coda (optional). The optional onset and coda represents the sounds of a syllable at its boundaries, which accounts for the cross-syllable coarticulation effects. To capture the coarticulation, the onset and coda are represented with distinctive features shown in Table 3.1 and 3.2 respectively for Cantonese and Mandarin. Distinctive features are minimal linguistic units which distinguish maximally close phonemes.

Cantonese	
Onset	*Coda*
Alveolar	Alveolar
Glide	Labial
Neutral	Velar
Labial	—
Lateral	—
Velar	—

Table 3.1: Distinctive features used to represent onset and coda of Cantonese syllables.

The distinctive features describe the place of articulation in speech production. For example, LABIAL refers to using the lips, ALVEOLAR refers to placing the tongue at the alveolar ridge (in pronouncing /d/, /t/, /s/, /z/, etc.), and VELAR refers to raising the velum which separates the nasal and

Mandarin	
Onset	*Coda*
Alveolar	Alveolar
Velar	Velar
Labial	Mid-Front
Lateral	Retroflex
Neutral	Rounded
Glide	Glide
Retroflex	Back
Palatal	—

Table 3.2: Distinctive features used to represent onset and coda for Mandarin syllables.

oral cavities (in pronouncing /ng/, /k/, etc.) Details are described in [2].

Syllable acoustics are affected by their neighboring contexts. For example, consider the character sequence "六七八" (i.e. six seven eight), pronounced as /luk cat baat/. The syllable for "七"(i.e. seven), /cat/, should end with an ALVEOLAR feature, but this feature is assimilated with the LABIAL onset of the syllable /baat/ (corresponding to "八"(i.e. eight)). Hence in this number sequence, the syllable /cat/ (for "七") is often produced as /cab/ (like the pronunciation of "輯") instead. If we ignore such coarticulatory effects, and concatenate this syllable for another number sequence, e.g. "八七六" (i.e. eight seven six), such that /cab/ is not followed by a LABIAL onset, the generated output will sound problematic.

Such considerations motivates our approach described as follows: each tonal syllable is augmented with a pair of digits which represent the distinctive features in the left and right contexts respectively. As illustrated in

Figure 3.1, the target tonal syllable (i.e. TSyl in the center of the triplet in Figure 3.1) has a digit (represented as "L") to encode the distinctive feature of the coda of its left neighbour, and a digit (represented as "R") to encode the distinctive feature of the onset of its right neighbour.

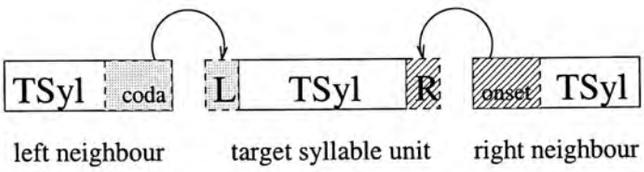


Figure 3.1: Two-digit encoding of the coarticulation with neighboring syllables at the syllable boundaries.

We chose to capture the coarticulation effects with distinctive features instead of phonemes in order to reduce the number of combinations of left and right contexts for each syllable. The representation of two digit encoding gave us a systematic way to name the syllable tokens as well as to search for a desired syllable token.

3.2 Creating a Domain-Optimized Wavebank

Our previous section shows that we need to store *multiple* tokens for each syllable in order to capture its coarticulatory variants. Hence we need to create a wavebank that contains all the syllable tokens that we will need for a given application domain. Our feasibility demonstration aims to generate a spoken delivery of FOREX information, e.g. the raw data [C, USD, HKD, 7.7743, 7.7744] is verbalized in Cantonese as "二零零一年六月一日，上

午八點十二分，歡迎使用外幣對換價既查詢服務。你需要既外幣匯價係，美元匯價對港元，買入七點七七四三，賣出七點七七四四。多謝你使用我地既服務，拜拜。". The approximate translations is "June first two-thousand and one, eight twelve a.m., welcome to our FOREX inquiry system. The foreign exchange rate you requested is, US dollar (USD) to Hong Kong dollar (HKD), bid seven point seven seven four three, ask seven point seven seven four four. Thank you for using our system. Goodbye. ". The underlined are information provided by the raw data input.

Conversion from the raw data to speech output involves appropriate verbalization followed by pronunciation dictionary lookup. Verbalization is critical since a specific datum, e.g. 12, should be verbalized as "十二月" if it refers to a month, but as "十二" if it is part of an exchange rate before the decimal point, and as "一二" if it is part of an exchange rate after the decimal point. We use a grammar to encode such heuristics and semantics needed for proper verbalization. Figure 3.2 shows an excerpt of our grammar for the major information categories in the FOREX domain. This grammar defines the overall sentential structure of the generated speech in FOREX, it also defines the coverage of the left and right coarticulatory contexts needed to be considered. This sentential structure implicitly defines the prosodic structure of our generated speech. Each information category is represented in sub-grammars, as illustrated in Figure 3.3.

The verbalized text should subsequently be converted into the appropriate syllable sequence by pronunciation dictionary lookup. For example, in

date *time*，歡迎使用外幣對換價既查詢服務。*currency pairs to buy/sell*。 *bid/ask rates*。多謝你使用我地既服務，拜拜。

Figure 3.2: High level grammar for FOREX response. Sub-grammars are underlined. The definition of the sub-grammars are in Figure 3.3.

date: [*year*][*month*][*day*]
time: [*am/pm*][*hour*][*minute*]
currency pairs to buy/sell: 你需要既外幣匯價係，[*selected_currency*][*base_currency*]。
bid/ask rates: 買入 [*bid_rate*]，賣出 [*ask_rate*]。

Figure 3.3: Sub-grammars for various information categories in FOREX response generation.

Mandarin, the verbalized form of the bid rate "7.7743" (i.e. "七點七七四三") is pronounced as /qi dian qi qi si san/ and if we include the left and right codes for distinctive features, this sequence becomes: /0_qi_1 4_dian_6 1_qi_6 4_qi_1 4_si_1 4_san_0/. In Cantonese, The same bid rate is pronounced as: /cat dim cat cat sei saam/, the encoded syllable sequence is /0_cat_2 2_dim_2 1_cat_2 2_cat_9 2_sei_2 9_saam_0/.

In order to create a compact wavebank that contains all the possible syllable tokens needed for the generation of possible outputs specified by our grammar, we have designed the *generate-and-filter* algorithm. We came up with this algorithm based on the idea that we wanted to cover only the significant sentences (which contain as many desired syllables as possible) as our recording prompts. The detail of the generate-and-filter is described in the following subsection.

3.2.1 Generate-and-Filter

A *generate-and-filter* algorithm is used to produce a compact set of recording prompts. First the response grammar is used to generate the possible response expressions. This is referred to as our generated set. It includes carrier phrases with the appropriate prosodics for our application domain. These sentences are transformed into syllable-based units by looking up their pronunciations from dictionaries. The five steps of the filtering process is listed below and the algorithm flow is illustrated in Figure 3.4.

Language	Generated Set	450 sentences
Cantonese	3860 sentences	450 sentences
Mandarin	4870 sentences	650 sentences

Table 3.3: Number of sentences in the generated set and filtered set.

Step 1 Compile the set of distinct acoustic units (with the two-digit context encoding) from the generated set.

Step 2 Compute a score for each acoustic unit. Each unit score is the inverse of its number of occurrences in the generated set.

Step 3 Compute the score of each sentence. Each sentence score is the summation of the acoustic unit scores in the sentence. If all sentences score zero, the process end.

Step 4 Sort the sentences by their scores. Move the highest scoring sentence from the generated set into the filtered set.

Step 5 Reset all scores in the generated set to zero. Goto Step 1

This subsequent filtering process aims to compress the generated set but retain all the contextual variations of the existing acoustic units. This compressed set becomes our filtered set. Table 3.3 shows the size of our generated and filtered set of sentences. The result shows that the generate-and-filter algorithm is able to compress roughly by a factor of 8. The resultant filtered set is of a manageable size for recording.

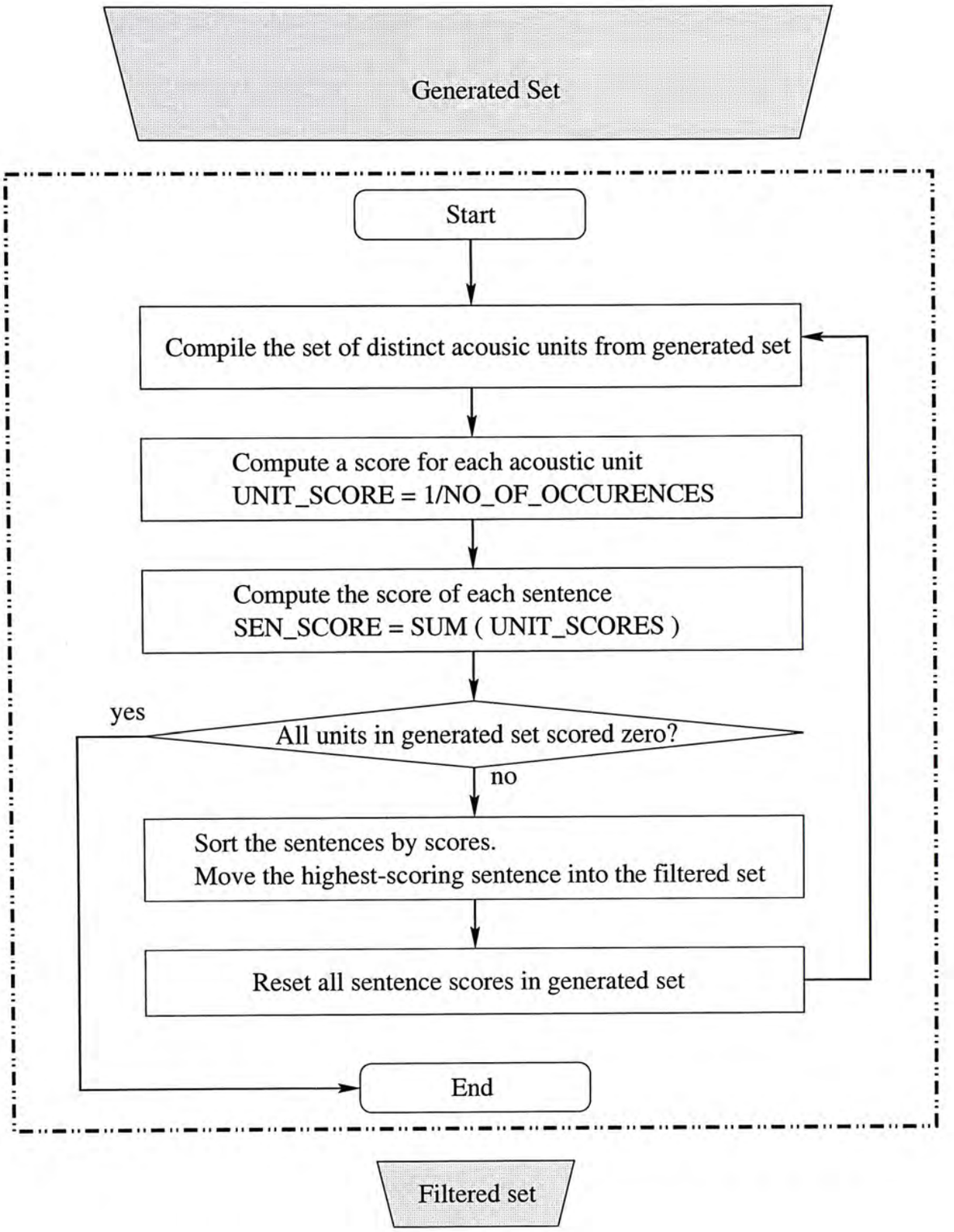


Figure 3.4: Flow of the generate-and-filter algorithm.

Given that we have produced a compact set of recording prompts that covers all coarticulatory context in the FOREX domain, we proceed to find a voice talent to record the waveforms and segment them into syllables. Only a single tonal syllable token is stored for each left and right context.

3.2.2 Waveform Segmentation

The segmentation of the recorded sentences is carried out manually using spectrograms. We have also begun to use a syllable recognizer to provide a forced alignment as an initial segmentation, to be hand-refined as a next step. The recognizer is trained with a set of 20,000 sentences, for the Hidden Markov Model of syllable initials and finals. Then we perform forced alignment between the waveform and the transcribed syllables to segment the waveforms into a sequence of syllable-based units.

Figure 3.5 shows a spectrogram of the Cantonese sentence "七千八百三十一點八三一" (i.e. seven thousand eight hundred and thirty one point eight three one) pronounced as /cat cin baat baak saam sap jat dim baat saam jat/. Figure 3.6 shows the spectrogram with alignments on syllable boundaries.

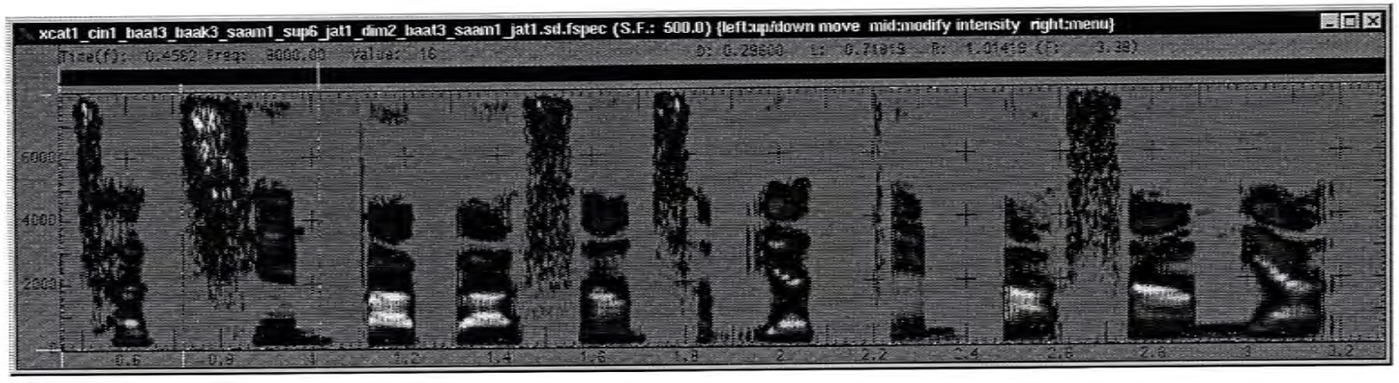


Figure 3.5: Spectrogram of the sentence "七千八百三十一點八三一" recorded in Cantonese.

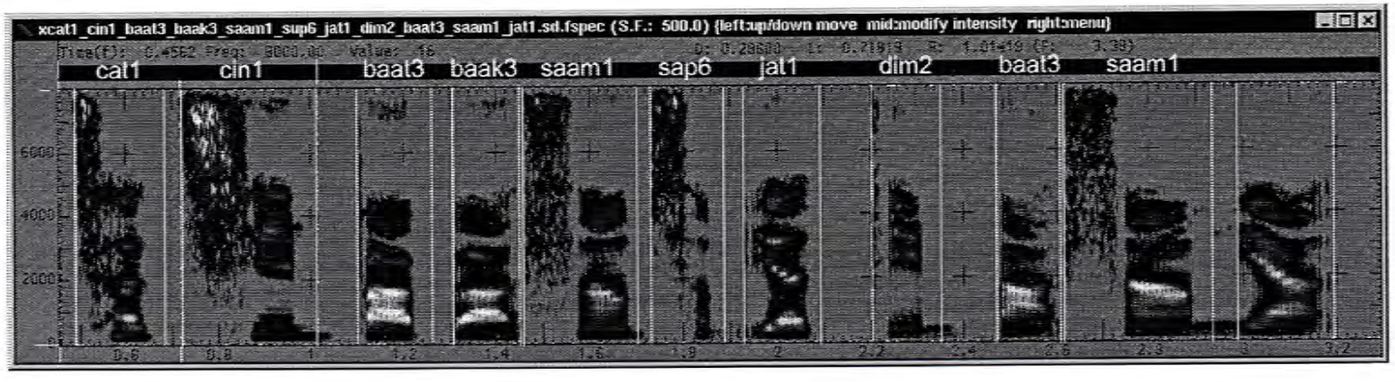


Figure 3.6: Spectrogram in Figure 3.5 aligned on syllable boundaries. The alignments are indicated with white lines.

Precise segmentation is important for the quality of the synthesis outputs. We recorded from two female speakers, one for Cantonese and the other for Mandarin. The process of segmentation yields 2,400 acoustic segments. Each segments contains a syllable or a multi-syllable unit which is described next.

3.3 The Use of Multi-Syllable Units

Our approach for domain-optimized speech generation also involve the use of units larger than the syllable. These contain fixed, invariant phrases, such as "歡迎使用外幣對換價既查詢服務" (Translation: Welcome to our FOREX inquiry system); and frequently occurring words that characterize our domain. For example, in FOREX, each name of the thirty some globally traded currencies is stored as a single waveform for concatenation. We also store the coarticulatory variants of these currency names as well. Consider, the phrase:

"港元匯價對美元, 買入七點七七四三"

(Translation: Hong Kong dollar to US dollar, bid seven point seven seven four three)

may be generated by concatenating 15 syllables as:

/0_gong_4/ /3_jyun_9/ /2_wui_3/ /9_gaa_2/ /9_deoi_1/ /9_mei_4/ /9_jyun_1/, /2_maai_4/ /9_jap_2/ /1_cat_2/ /2_dim_2/ /1_cat_2/ /2_cat_2/ /2_sei_2/ /9_saam_0/

but we choose to generate it as:

/0_gong-jyun-wui-gaa_2/ /9_deoi-mei-jyun_1/, /2_maai-jap_2/
/1_cat_2/ /2_dim_2/ /1_cat_2/ /2_cat_2/ /2_sei_2/ /9_saam_0/,

The multi-syllable units used are boldfaced. They refer to, in order, "Hong Kong dollar", "US dollar" and "bid". This is because we can minimize the number of concatenations and potential distortions involved to achieve a better speech quality. It can be seen that the multi-syllable units are treated the same way as a single syllable unit in our framework.

3.4 Unit Selection for Concatenative Speech Output

Having created a domain-optimized wavebank, we should be able to generate a spoken delivery of some raw data on demand. This section describes our approach in selecting the appropriate (multi-)syllable units for concatenation.

Our unit selection process is quite simple. For a given textual input which is mapped into syllable-based units, our synthesis algorithm concatenates the corresponding acoustic wave files sequentially from left to right. The unit selection process ensures that the acoustic segments with matching left and right contexts are chosen.

In addition, tone sandhi rules are applied in Mandarin synthesis. If there is a series of syllables with the third tone, all the syllables are changed to the

second tone except for the last one. For example, the phrase 九 /gau3/ (tone 3) 九 /gau3/ (tone 3) 九 /gau3/ (tone 3) (i.e. nine nine nine) is read as 九 /gau2/ (tone 2) 九 /gau2/ (tone 2) 九 /gau3/ (tone 3). Short pauses are also inserted in between phrases, and long pauses in between sentences. Both the unit selection process and the insertion of pauses were found to be important contributing factors towards naturalness in the synthesized outputs.

3.5 A Listening Test

We have designed a listening test to assess the effectiveness of domain-optimization. Our domain-optimized speech generation output is compared against a Cantonese text-to-speech (TTS) system [7] that involves no domain-optimization. This synthesizer can handle domain-independent textual input by TD-PSOLA synthesis. In comparison, our current synthesis task is simpler and more restrictive due to domain-specificity. We hope to show that the effort devoted to optimization within the domain contributes towards higher intelligibility and naturalness of the synthesized outputs.

We set up the experiment as follows: A listening test is set up as a within-group experiment involving 12 subjects. Ten pairs of synthesis outputs are generated from ten sentences, that is, two waveforms per sentence. The sentences cover all the currencies within our foreign exchange domain, and their exchange rates at various dates and times. One of the waveform pairs is generated by the TD-PSOLA synthesizer, and the other by the current syllable

concatenation technique. The order of the waveforms are randomized to neutralize learning effects. Each subject is asked to rate the pair of waveforms in terms of intelligibility and naturalness, on a scale of 1 to 6 (1 represents barely intelligible / natural, and 6 represents extremely intelligible / natural). Instructions and demonstrations about naturalness and intelligibility are given to the subjects, to ensure that they have enough knowledge for the judgment. The questionnaire is shown in Appendix A.

We formulated a *t-test* using the difference in opinion score as our test statistic. The differences in intelligibility scores have a mean of 1.2 and a standard deviation of 1.17. The differences in naturalness scores have a mean of 1.8 and a standard deviation of 1.15. Testing at a significance level of 0.05 concludes that we should accept the alternate hypothesis. That is, syllable concatenation is more intelligible and natural than TD-PSOLA within the foreign exchange application. The details of the statistical test is shown in Figure 3.7.

3.6 Chapter Summary

In this chapter we have presented our approach for domain-optimized Chinese speech generation by the technique of syllable concatenation. Syllable coarticulation is captured by the use of distinctive features. We have designed a generate-and-filter algorithm which helps create a domain-optimized wavebank that is compact yet contains all the possible syllable coarticulatory

The parameter of interest is the difference in **intelligibility** score μ_I
$H_0 : \mu_I = 0$
$H_1 : \mu_I > 0$
$\alpha = 0.05$
The test statistic is
$t_0 = \frac{\overline{x} - \mu_{I0}}{s/\sqrt{n}}$
Reject H_0 if $t_0 > t_{0.05,11} = 1.796$
Computation: $\overline{x} = 1.2$, $s = 1.17$, $\mu_{I0} = 0$, and $n = 12$, we have
$t_0 = \frac{1.2-0}{1.17/\sqrt{12}} = 3.553$
Reject H_0 if $t_0 > t_{0.05,11} = 1.796$
Conclusion: Since $t_0 = 3.553 > 1.796$
we reject H_0 and conclude at the 0.05 level of significance that the syllable concatenation is more **intelligible** than TD-PSOLA within the FOREX domain

The parameter of interest is the difference in **naturalness** score μ_N
$H_0 : \mu_N = 0$
$H_1 : \mu_N > 0$
$\alpha = 0.05$
The test statistic is
$t_0 = \frac{\overline{x} - \mu_{N0}}{s/\sqrt{n}}$
Reject H_0 if $t_0 > t_{0.05,11} = 1.796$
Computation: $\overline{x} = 1.8$, $s = 1.15$, $\mu_{I0} = 0$, and $n = 12$, we have
$t_0 = \frac{1.8-0}{1.15/\sqrt{12}} = 5.422$
Reject H_0 if $t_0 > t_{0.05,11} = 1.796$
Conclusion: Since $t_0 = 5.422 > 1.796$
we reject H_0 and conclude at the 0.05 level of significance that the syllable concatenation is more **natural** than TD-PSOLA within the FOREX domain

Figure 3.7: Details of statistical testing on listening test data, regarding the intelligibility and naturalness of syllable concatenation against TD-PSOLA.

variants that may occur within the domain. We have also defined the criteria for selecting the "appropriate" syllable variant during the generation process. This selection criteria enforce matching left and right coarticulatory contexts in choosing the syllable for concatenation. A listening test shows that our approach compares favorably against a non-domain-optimized Cantonese TTS system for the FOREX domain.

Chapter 4

Scalability and Portability to the Stocks Domain

In the previous chapter, we described the feasibility of the speech synthesis framework in a relatively constrained domain. In order to test the scalability and portability of our approach, we migrated our technique of corpus-based syllable concatenative synthesis from the foreign exchange (FOREX) domain to the stocks domain for a system known as ISIS (Intelligent Speech for Information Systems) [4, 5]. This chapter presents the procedure in porting our speech generation framework from the FOREX domain to the ISIS domain. We also enhanced our framework for scalability since ISIS is much more complex than FOREX.

4.1 Complexity of the ISIS Responses

Our speech generation technique is intended for Chinese spoken response generation in the ISIS spoken dialog system. We generate both Cantonese and Mandarin in ISIS. This domain covers user requests and inquiries in nine categories:

1. *real-time stock quotes*

   E.g. "0005匯豐控股成交價是一百元。" (Translation: The current price of 0005 Hong Kong and Shanghai Bank is one hundred dollars per share.)

2. *securities trading*

   E.g. "請確認你的指示: 買入0005匯豐控股四百股，每股一百元。你現在要執行這個指示嗎?" (Translation: Please verify your request: purchase four hundred of 0005 Hong Kong and Shanghai Bank for one hundred dollars per shares Would you like to place this order?)

3. *order status enquiries*

   E.g. "你於二零零一年八月廿四日十一時三十分買入0005匯豐控股四百股，每股一百元。" (Translation: You have purchased four hundred shares of 0005 Hong Kong and Shanghai Bank for one hundred dollars per share at eleven thirty a.m., on Tuesday the twenty-forth of August, two thousand and one.)

4. *buy/sell order amendments*

E.g. "請確認你的指示: 更改買入0005匯豐控股四百股，每股一百元。新指示爲買入0005匯豐控股三百股，每股一百元。你現在要執行這個指示嗎?" (Translation: Please verify your request: amend a former purchase order of four hundred shares of 0005 Hong Kong and Shanghai Bank for one hundred dollars per share, to a new purchase order of three hundred shares of 0005 Hong Kong and Shanghai Bank for one hundred dollars per share. Would you like to proceed with the processing?)

5. *buy/sell order cancellations*

   E.g. "請確認你的指示: 取消賣出0005匯豐控股四百股，每股一百元。你現在要執行這個指示嗎? " (Translation: Please verify your request: cancel a former sell order of four hundred shares of 0005 Hong Kong and Shanghai Bank for one hundred dollars per share. Would you like to place this order?)

6. *market trends*

   E.g. "現在恆生指數是一萬二千點。" (Translation: Currently the Hang Sang Index is twelve thousand points.)

7. *portfolio/account information*

   E.g. "你持有0005匯豐控股一千股，每股盈利五元。" (Translation: You are holding one thousand shares of 0005 Hong Kong and Shanghai Bank, purchased at fifty HK dollars per share. It is now gaining five dollars per share.)

8. *financial news*

   E.g. "對不起，0005匯豐控股無新聞" (Translation: Sorry, there's no news about 0005 Hong Kong and Shanghai Bank. Please take a look.)

9. *chart display.*

   E.g. "0005匯豐控股的最新走勢圖，請看。" (Translation: Here's the latest trend for 0005 Hong Kong and Shanghai Bank. Please take a look. (Display graph on screen))

More examples of the nine major categories in Chinese and Mandarin are shown in Appendix B.

These nine categories of responses require 22 sentential grammar rules for generation (some categories require multiple sentential grammar rules). In addition, the ISIS spoken dialog calls for generation of error messages and system confirmation message. An example of such an error message is: "對唔住，我唔明白你既指示" (i.e. Sorry, I do not understand your instruction.). Error/confirmation messages another 34 sentential grammar rules. Consequently the response grammar contains 56 sentential grammar rules in total. Each rule may invoke sub-grammars defining *date*, *time*, etc. Recall that in FOREX, we only have a single sentential grammar rule for generation. Hence in comparison, ISIS is substantially more complex than FOREX. To address the issues of portability and scalability from FOREX to ISIS, we have incorporated several enhancements in our Chinese speech generation framework:

1. XML for raw data input semantic and sentential grammar specification – We adopt the XML convention in tagging the semantics of raw data. We also use XML tags in the sentential grammar rules as well as sub-grammars. Following these XML tags, we can appropriate merge the semantics according to the grammar structure for generation. This enhances the reconfigurability and reusability of our response generation procedure.

2. Enhanced *generate-and-filter* algorithm – In order to create a domain-optimized recording corpus with a minimal size but maximal coverage of the syllable variants, we have described the *generate-and-filter* algorithm in the previous chapter. This algorithm includes the first step of exhaustive text generation of possible responses, followed by the second step of filtering to compress into a minimal set of recording prompts. However, the first step becomes formidable when the domain complexity increases, because exhaustive generation leads to combinatorial explosion. In this chapter we present our enhanced generate-and-filter algorithm which tightly couples generation and filtering in a tree-based representation.

3. Energy normalization – Energy fluctuations among our syllable segments affect the overall synthesis quality especially when loud syllables are concatenated with soft syllables. This chapter also presents a simple enhancement where we normalize the energy or intensity over an

entire set of recordings prior to waveform segmentation. We hope to get syllable segments with more even intensity to be stored in our wavebank.

We will describe each of the three enhancements in detail as follows.

4.2 XML for input semantic and grammar representation

As we move from the FOREX domain to the ISIS domain, we need to deal with greater variety of semantics. This motivate us to use XML to tag the raw data with a appropriate semantic labels.

Consider the example of input raw data in the ISIS domain:

[C, 2001, 6, 12, 0005.hk, up, 3.4]

The raw data represents, in order, the language for generation(Cantonese), year (2001), month (6), date (12), ric_code (0005.hk), movement (up) and changes (3.4). With the use of XML tags, we have a more structural semantic frame as shown in Figure 4.1.

The boldfaced tags in Figure 4.1, i.e. < **response** >< **/response** >, < **language** >< **/language** > and < **grammar** >< **/grammar** > specify, respectively, the beginning of a response, the language or dialect for generation and the name of the response grammar for generation.

< **response** >
< **language** > *cantonese* < /**language** >
< **grammar** > *categoryV*3 < /**grammar** >
< *today* >< *yy* > 2001 < /*yy* >< *mm* > 6*m* < /*mm* >< *dd* > 12*d* < /*dd* >< /*today* >
< *ric_code* > 0005.*HK* < /*ric_code* >
< *movement* > *up* < /*movement* >
< *change* > 3.4 < /*change* >
< /**response** >

Figure 4.1: Example of an input semantic frame. This specifies the response should be generated in Cantonese (< *language* >), with the third grammar in the response type "Real-time Quote".

The remaining tags are semantic in nature, e.g. < *ric_code* >< /*ric_code* > stands for the Reuters Instrument Code for a stock; < *movement* >< /*movement* > stands for the increment or decrement in price and < *change* >< /*change* > stands for the relative change in price.

XML tags are also used to label response grammar rules. Consider the sentential grammar rule as shown in Figure 4.2:

< *tmpt name* = *date* > *today* < /*tmpt* >
< *option* > *ric_code* < /*option* >
< *pause* >
< *fix* > *each* < /*fix* >
< *option* > *movement* < /*option* >
< *price* > *change* < /*price* >

Figure 4.2: Example of grammar rule.

This sentential grammar characterize generated responses such as "零零零五匯豐控股(ric_code), 每股(each) 上升(movement) 四毫(change)" (Note

that the semantics in this response are in the same order as the tags appear in the XML-tagged grammar rule of Figure 4.2). The XML tags in the grammar rule of Figure 4.2 control the generation process(es) of the response, described as follows:

1. $< fix >< /fix >$ – The tag $< fix >$ specifies a static syllable (or multi-syllable) segment. This is a pre-recorded segment, labeled as *each* (see Figure 4.2 in the wavebank, and will be retrieved directly for further response generation.

2. $<$ *option* $><$ */option* $>$ – The tag $<$ *option* $>$ specifies that the datum labeled with the semantic tag *movement* should be fetched from the input semantic frame (see Figure 4.1. The datum extracted will be mapped into its Chinese verbalization, which is retrieved from the wavebank for response generation. In this example, "$<$ *option* $>$ *ric_code* $<$ */option* $>$" refers to the RIC code from the semantic frame, whose value is 0005.hk. The multi-syllable segment labeled as 0005.hk reads "零零零五匯豐控股".

3. $<$ *price* $><$ */price* $>$ – The grammar tag $<$ *price* $>$ specifies a datum to be fetch from the input semantic frame. This additionally undergoes a verbalization process, with is customized for numerical price expressions. In this example, the grammar tag $<$ *price* $>$ refers to the verbalization of the datum 3.4 in the semantic frame to "三蚊四毫". The verbalized form is then mapped into a syllable sequence that

govern concatenative generation.

4. < *number* >< /*number* > – There is also a grammar tag < *number* > that operates in a way similar to < *price* >. The difference lies in the verbalization process for general numeric expressions and not price expressions. E.g. if the datum 3.4 is tagged as < *number* >, it will be verbalized as "三點四".

5. < *tmpt name* = TEMPLATE > – The grammar tag < *tmpt name* = TEMPLATE > is used to call a sub-grammar rule, where TEMPLATE is the name of the sub-grammar. In our example, the template invokes the sub-grammar "date" which generates the expression: "二零零一年六月十二日" (i.e. June twelve, two thousand and one).

4.3 Tree-Based Filtering Algorithm

Compare with the FOREX domain, ISIS has a much greater variety in responses to be generated. We cannot create a wavebank for ISIS by the generate-and-filter algorithm (described in Section 3.2.1) because the algorithm involves an overly large expansion of recording prompts. Instead we design a tree-based filtering approach that can perform generation and filtering at the same time. This also shortens the time needed for creating the wavebank.

The tree-based filtering algorithm is an enhanced version of the generate-

and-filter algorithm, and can be described in the steps below:

Initialization Starting with level 0 (tree-root)

Step 1 Expand the node in current level with the vocabulary in next level.

Step 2 Examine each node on the new level one by one with its syllable and distinctive feature of its left context. If such combination (referred to as "syllable-feature combination" below) has occurred before, the node stops expanding. Else, expand the node and go to step 1.

The tree-based algorithm is illustrated in the Figure 4.3. The example is related to the response grammar for securities trading (please see Figure 4.4). This grammar can generate sentences such as "請確認你既指示, 買入零零零五匯豐控股一百股, 每股九十蚊. 你而家係唔係要執行呢個指示呢?" (partially shown in Figure 4.3).

The root (level 0) of the tree (shown in Figure 4.3) for this response grammar is a multi-syllable segment of "please_confirm". It expands with the vocabulary of the next level (level 1) as its children, i.e. "買入" (bid) and "賣出" (ask). The two nodes are examined with its "syllable-feature combination". The "syllable-feature combination" for "買入" (bid) is "(default)-/maai5-jap6/", for "賣出" (ask) is "(default)-/maai6-ceot1/)" (Table 4.1 shows the distinctive features for syllable coda). In this case, their "syllable-feature combinations" are unique (differ by their syllable). Therefore, the two nodes both expand with the vocabulary of level 2 (stocknames).

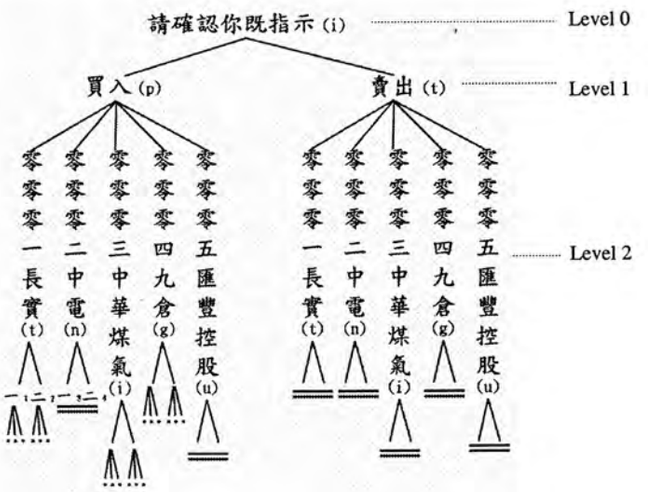


Figure 4.3: Example of tree-based filtering algorithm for the response grammar for "Securities Trading".

< *fix* >please_confirm< /*fix* > (LEVEL 0)
< *pause* >
< *option* >buysell< /*option* > (LEVEL 1)
< *option* >stockname< /*option* > (LEVEL 2)
< *tmpt name* = *share* > *buyshare* < /*tmpt* >
< *pause* >
< *fix* >each< /*fix* >
< *price* >shareprice< /*price* >
< *pause* >
< *fix* >request< /*fix* >

Figure 4.4: Response grammar for "Securities Trading". The texts in parentheses shows the corresponding tree level in Figure 4.3.

Articulatory Characteristic	Syllable Coda
Labial	p, m
Alveolar	t, n
Velar	g, k
Default	others

Table 4.1: Articulatory characteristic of the left context (syllable coda) represented in distinctive features.

Each node at level 2 (e.g. "零零零一長實") is examined. As all nodes in level 2 have different "syllable-feature combinations" (either differ by their syllables or by features of their left context), they all expand to level 3. The nodes at level 3 are only partially illustrated in Figure 4.3. The nodes are numeric digits (of stock shares). Those nodes with parent "零零零二中電" stop expanding (denoted with two parallel lines in the Figure) because they have identical "syllable-feature combinations" with the nodes expanded from the node "零零零一長實" (There is a small digit mark at the right bottom of the nodes "一" and "二". Nodes with mark 1 and 3 has combination of "(alveolar)-/jat1/"; nodes 2 and 4 have combination of "(alveolar)-/ji6/"). The tree is expanded accordingly level by level until it reaches the end of the grammar. Finally, all the possible paths from root to leaf of the tree form the set of recording prompt.

We compare the speeds of the original generate-and-filter and the new tree-based filter algorithms in wavebank creation, based on generating seven-digit numbers. Seven-digit numbers can be used to represent the bid/ask rates in FOREX domain. The grammar for generation is

[*digit*]千[*digit*]百[*digit*]十[*digit*]點[*digit*][*digit*][*digit*]

[*digit*] represents Chinese number one("一") to nine("九").

With the generate-and-filter algorithm, we generated 3.64Kb of Chinese text for recording prompts in 12.5 minutes. With the tree-based filtering algorithm, we generated 10.1Kb of Chinese text for recording prompts in 0.13 seconds. The two sets of recording prompts are for the same seven-digit grammar. Notice that the tree-based filtering algorithm has a much faster speed in generating the same set of recording. Nevertheless, the resulting set of recording prompts is less compact. It is because the generate-and-filter algorithm obtains a desired syllable units in a sentence that is generated at the syllable's first occurrence: the sentence may contain a lot of other undesirable syllables. Whereas, the generate-and-filter algorithm ensures that a desired syllable is obtained from a sentence that also contains maximum number of other desired syllables.

4.4 Energy Normalization

Energy normalization is a technique applied to adjust the energy of the recordings to a more consistent level across syllable segments. It helps to improve the speech quality of Chinese speech generation.

Energy normalization involves Equations (4.1) to (4.3). An ideal unit energy N_{ideal} is first set for an appropriate volume with formula (4.1). Unit

energy refers to the total energy per unit time.

$$N_{ideal} = \frac{\sum_{i=1}^{n} X_i^2}{d} \tag{4.1}$$

It is done by selecting an acoustic segment which has a desired volume, then X_i is the signal sample magnitude and d is the duration of the acoustic segment. The total energy of the selected acoustic segment is computed and normalized with its duration. The ideal unit energy N_{ideal} is thus yielded.

To adjust the energy of a recording with duration d and signal sample x_i with reference to N_{ideal}, we applied Equation (4.2). to obtain a multiplicative factor f. This factor will be used to amplify or de-amplify its corresponding recordings. In the way, we even out the energies of all our recordings.

$$\frac{\sum_{i=1}^{n} (f \times x_i)^2}{d} = N_{ideal} \tag{4.2}$$

$$f = \sqrt{\frac{N_{ideal} \times d}{\sum_{i=1}^{n} x_i^2}} \tag{4.3}$$

The factor f is obtained by multiplying the ideal unit energy N_{ideal} with its duration d, and normalized on its total energy. Afterwards, the acoustic signals are amplified or de-amplified with this factor. After energy normalization, the volume of acoustic segments are more consistent. The result of energy normalization is illustrated in Figure 4.5. The figure shows that the energy level of the concatenated syllables are less fluctuated (indicated by

the dotted line).

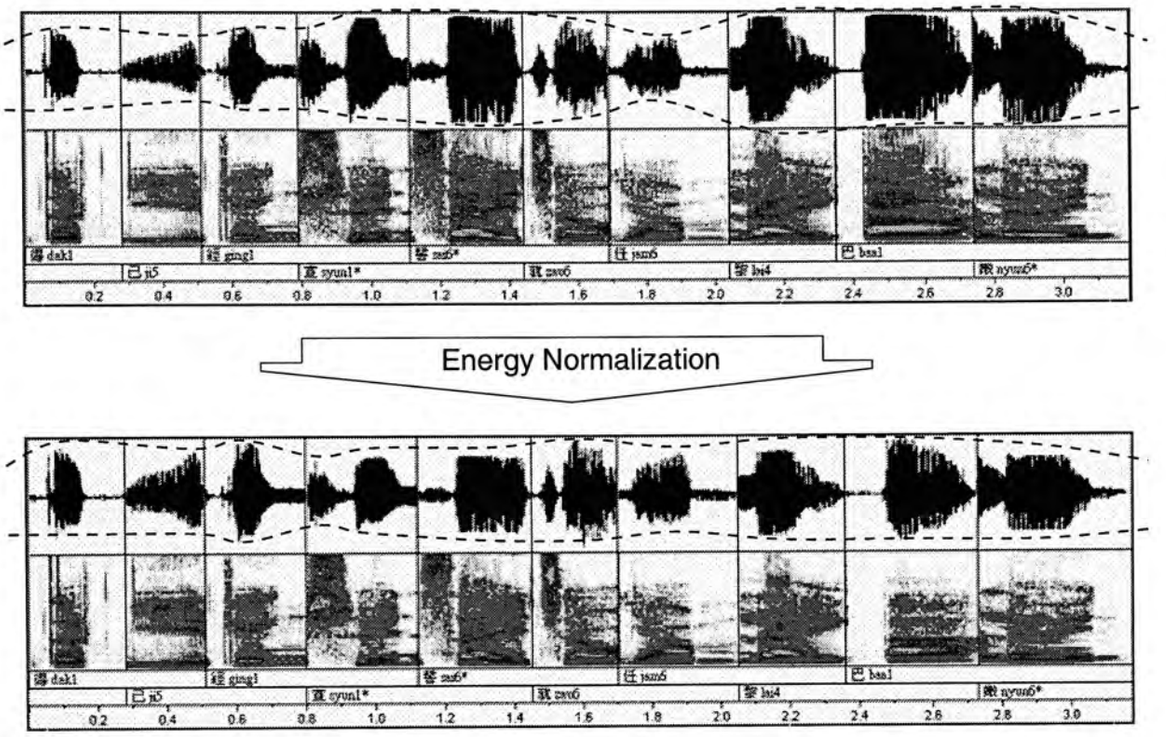


Figure 4.5: Result of energy normalization – notice the energy fluctuations indicated by the dotted lines are reduced after energy normalization.

4.5 Chapter Summary

This chapter describes the scalability and portability of our speech generation framework to a more complex domain. It is done by plugging the framework to a spoken dialog system of the stocks domain, in which 56 sentential grammar rules are required to model various response types. We have made three major enhancements related to framework architecture and wavebank creation: XML is adopted for the input semantic frame and the grammar rules;

the tree-based filtering algorithm is developed for large scale wavebank construction; and the technique of energy normalization is applied to adjust the energy level of the recordings to a more consistent level.

Chapter 5

Investigation in Tonal Contexts

In our approach for speech generation by corpus-based syllable concatenation, we have thus far modeled only the coarticulatory effects due to the place of articulation. However, for a tonal language such as Chinese, correct pronunciation of the syllable tone is very important as well. The realization of a syllable's tone (in terms of the fundamental frequency or pitch contour), is also affected by the tones of the neighboring syllables.

In the previous applications of FOREX and ISIS, since the generation grammar is constrained within one to a few carrier phrases, the effect of tonal context was not too pronounced. However, as we scale up to more complex domains, there may be more tonal context variations, which imply that in addition to the consideration in place of articulation, we must consider tonal context as well. An example is shown in Figure 5.1 to illustrate the tonal distortion that may occur upon concatenation when tonal context is

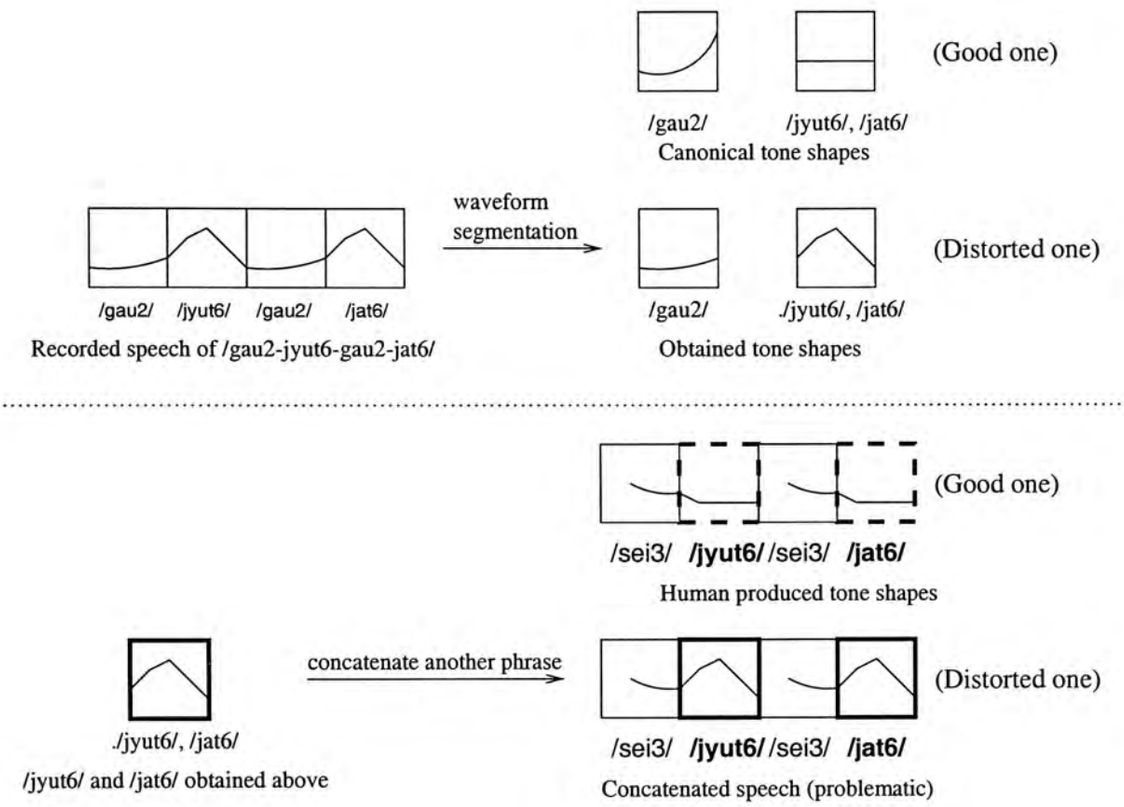


Figure 5.1: Example of tonal distortion upon concatenation when tonal context is not considered.

not considered.

In Figure 5.1, the tone shapes of the syllables /gau2/, /jyut6/ and /jat6/ (corresponding to "九", "月" and "日" respectively) segmented from the recording of "九月九日" (i.e. the ninth of September), whose pronunciation is /gau2 jyut6 gau2 jat6/, is distorted from their canonical tone shapes. If the syllables, /jyut6/ ("月") and /jat6/ ("日") are concatenated to generate, for example, "四月四日" (i.e. the fourth of April), whose pronunciation is /sei3 jyut6 sei3 jat6/ the generated output will sound problematic, because the pitch contour of the concatenated speech is very different from what it

should be in human produced speech.

The scope of our investigation in tonal context is focused on Cantonese. There are two major goals in our investigation.

1. To find out the relative importance between the left and right tonal contexts. This part of the investigation involves two subgrammars – the DATE-TIME subgrammar and the NUMERIC subgrammar. The DATE-TIME subgrammar involves fewer variations in tonal context for comparing the left with the right contexts. The NUMERIC subgrammar involves all combinations of number triplets in Cantonese. The Cantonese numbers cover all six tones, and thus provide more variations in tonal context than the DATE-TIME subgrammar. The results for these two subgrammars serve to corroborate our findings.

2. To establish a backoff scheme for unit selection. If we consider *both* the distinctive features (place of articulation) and neighboring tones in coarticulation, the number of possible coarticulatory contexts for each syllable increases very quickly. This implies we will need many more syllable variants in our wavebank to fully cover all possible contexts. However, we believe that full coverage of contextual variants may not be necessary. We hypothesize that certain tonal contexts may substitute for others without perceivable differences by the human listener. Therefore, this part of our investigation seeks to establish a "backoff scheme". When our concatenation algorithm calls for a syllable variant

that is absent from our wavebank, we can follow our backoff scheme to find the next best substitute for concatenation.

In this chapter we will first describe the nature of tones and their acoustic correlates. Then we will present our investigation in the relative importance between left and right tonal contexts, as well as the establishment of a backoff scheme for tonal variants during concatenation.

5.1 The Nature of Tones

Tone itself is a linguistic feature, its acoustic correlate is the fundamental frequency (f0). Speech contains a variety of signals with different frequencies, amplitudes and phases. The fundamental frequency (f0) is the lowest frequency among these signals.

There are three attributes in tone: tone height, tone shape and duration. We only focus on the first two attributes in our investigation. Tone height is the pitch level, tone shape describes the trajectory of the fundamental frequency within the syllable. Consider the tones in Cantonese. This Chinese dialect has nine lexical tones, which can be reduced to six if we only consider tone height and shape. This is illustrated in Figure 5.2. In this figure, we can observe different tone heights among the level tones 1, 3, 4 and 6. Tones 2 and 5 are called rising tones, as shown in the figure as well.

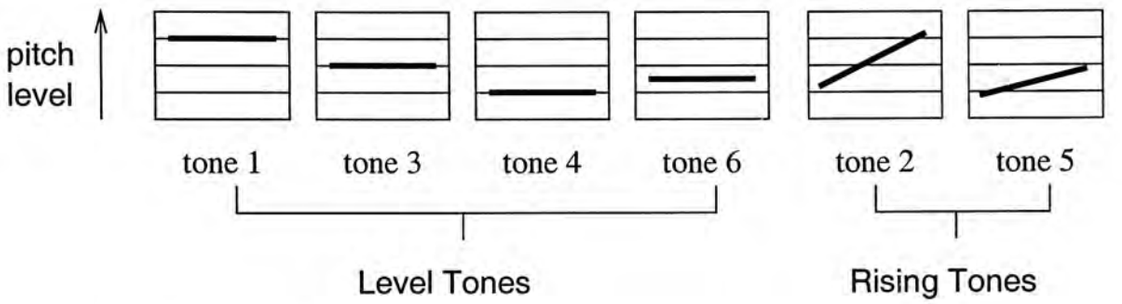


Figure 5.2: Cantonese tones categorized into two group based on their tone shapes.

5.1.1 Human Perception of Tones

The process of human perception of tones is illustrated in Figure 5.3. The process can be described in four steps:

Step 1: A speaker thinks of a syllable with a lexical tone

Step 2: The syllable's tone is articulated (encoded) in the form of fundamental frequency

Step 3: A listener receives the produced fundamental frequency

Step 4: The listener interprets (decodes) the fundamental frequency to a lexical tone

Produced tone –(encode)→ f0 –(decode)→ Perceived tone

Figure 5.3: Relation between produced tone and perceived tone.

The realization of a tone in terms of f0 is affected by the tones of the neighboring syllables, i.e. the tone shape may be distorted due to tonal context. The distortion produces variations in f0 and its trajectory, which may

affect correct decoding of the tone. Consider an example in continuous speech "零三六三上海實業" (i.e. 0363 Shanghai Industry Holdings) pronounced as /ling4 saam1 luk6 saam1 soeng6 hoi2 sat6 jip6/ (Figure 5.4).

Syllables such as /ling4/, /saam1/ and /jip6/ maintain their level tone shapes. However, the syllables /luk6/, /soeng6/ and /sat6/ that should be level now has rapidly falling shapes instead. This distortion is caused by the preceding syllables which end with a high pitch level, and the tone trajectory needs to rapidly move down to a low pitch level for /luk6/, /soeng6/ and /sat6/.

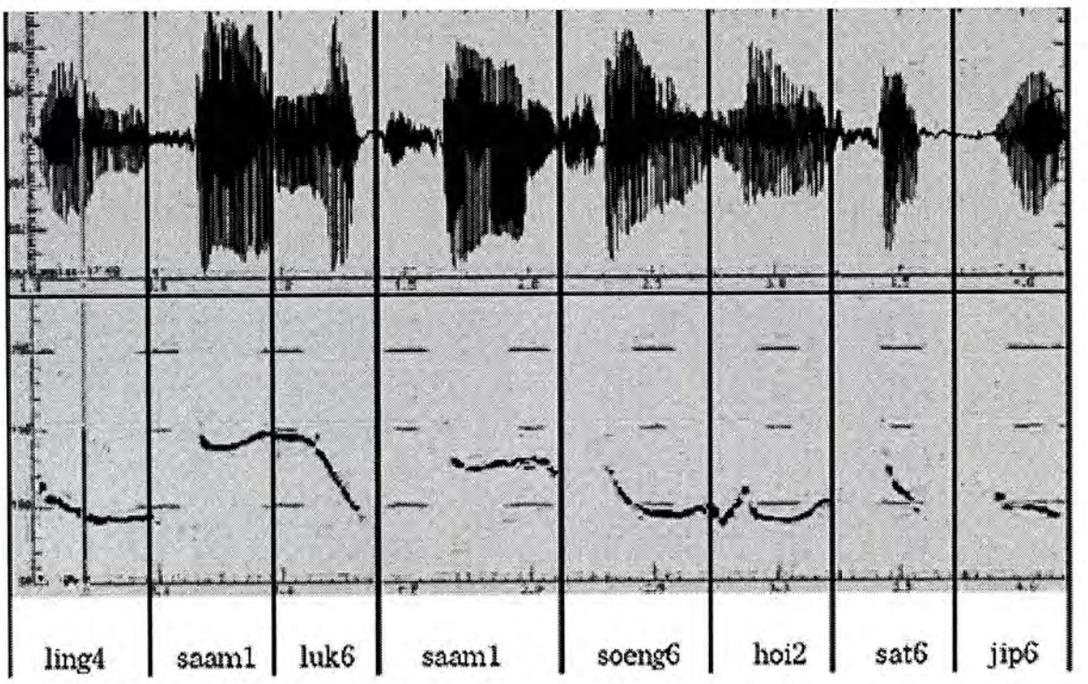


Figure 5.4: Example of the pitch contour in the real recorded phrase of "零三六三上海實業".

This falling tone shape is appropriate for the level tones such as /luk6/,

/soeng6/ and /sat6/ only for the appropriate tonal context shown. If we extract these tone shapes and insert in other tonal contexts, the non-level shape will lead to errors in perception (i.e. the human listener does not think that the syllable(s) have tone 6). These syllables with problematic tone shapes will probably perceived as tone 1 if they are concatenated with neighboring syllables in tone 4 (which has a low pitch level). This has critical implications for our speech generation approach. In order to achieve a high degree of naturalness, with correct tone perception in our speech outputs, we need to select syllable units with (coarticulated) tone shapes that fit the necessary tonal context.

We begin our investigation of tone by exploring the relative importance between the left tonal context and the right tonal context.

5.2 Relative Importance of Left and Right Tonal Context

5.2.1 Tonal Contexts in the Date-Time Subgrammar

We begin to compare the relative importance of left and right tonal contexts within the scope of the DATE-TIME subgrammar. This subgrammar generates outputs such as "二零零一年六月一日，下午三點二十分十一秒" (Translation: June first two thousand and one. three twenty and eleven seconds in the afternoon). The subgrammar contains two kinds of syllable

units: key units and value units. The key units mark the date(s) and time(s), e.g. "年", "月", "日", "上午/下午", "點", "分" and "秒". The value units mark the values of the date(s) and time(s) and are Chinese numbers such as "二零零一" and "六". One may think that since the key units are invariant, each key unit can be represented by a single syllable token during concatenation. In reality, the tone shape of each key unit's syllable is affected as its neighboring value units change. We chose this subgrammar as the scope of our investigation because it is relatively constrained and commonly used in many application domains. Also five out of the six tones in Cantonese are covered by the key units, as can be seen from the syllable pronunciations in Table 5.1.

Keys	**Meanings**	**Tonal Syllables**
年	year	nin4
月	month	jyut6
日	day	jat6
上午/下午	am/pm	soeng6-ng5/haa6-ng5
點	hour	dim2
分	minute	fan1
秒	second	miu5

Table 5.1: Key units of the date-time subgrammar and their corresponding tonal syllables.

In order to compare the effects of the left neighboring value with the right to see which is more important, we perform three tasks:

- Wavebank construction – We construct a wavebank that contains all tonal variants of all key and enough value units to create exhaustive

tonal context for the key units within the date-time grammar. A tonal variant is represented as $(L)TSyl(R)$, where $TSyl$ is the tonal syllable, L is the tone of the left syllable, and R is the tone of the right syllable.

We used the date-time subgrammar to exhaustively generate a set of recording prompts so as to cover all possible tonal contexts in terms of the value units. There are a total of 36 such recording prompts and an excerpt is provided in Figure 5.5 and the full list is given in Appendix C. In this figure, the value units are underlined, and the tonal context they provide for their neighboring key units is shown in parentheses.

二零零一(1)年(1)一(1)月(6)二(6)日，
下午(1)三(1)點(1)三十三(1)分(6)十一(1)秒
二零零三(1)年(2)九(2)月(6)六(6)日，
下午(1)一(1)點(1)三十九(2)分(6)十二(6)秒
二零零七(1)年(3)四(3)月(6)十(6)日，
上午(2)九(2)點(1)三十八(3)分(6)十三(1)秒
二零零一(1)年(5)五(5)月(6)二(6)日，
下午(3)四(3)點(1)三十五(5)分(6)十四(3)秒
二零零三(1)年(2)二(1)月(6)六(6)日，
下午(3)五(3)點(1)三十(5)分(6)十五(3)秒
.....

Figure 5.5: Example of the recording prompts for value units (underlined) to create various tonal environments for the key units.

If we consider the key units "日", "上午/下午" and "秒", their right tonal context is always NULL according to the subgrammar. Hence for each of them, we can extract six tokens for each tonal variant from our recording corpus in Appendix C Table C.1, e.g. you can find six in-

stances in Appendix C Table C.1 for the tonal variant consider "(1)秒". Only one of these six is store in our wavebank used later for concatenation. If we consider each of the remaining key units "年", "月", "點" and "分", most of them only have one instance for each tonal variant, e.g. there is only one instance of "(4)月(1)" in Appendix C Table C.1. Since we want to avoid the effect of segmenting a recording prompt and concatenating the original syllable segments to generate the *same* recording prompt, we designed *another* set of recording prompts to provide tonal variants for the key units. This set is given in Appendix C Table C.2 and C.3, and contains all tonal variants of all key units within the date-time subgrammar.

- Waveform generation – During generation by syllable concatenation, we should ideally enforce that the tonal variant chosen (i.e. $(L)TSyl(R)$) has a L value that agrees with the tone of the left syllable, and an R value that agrees with the tone of the right syllable. However, in order to compare the importance of the left tone context with the right, we also try to generate by selecting tonal variants with a mismatch in L and a match in R, or a mismatch in R and a match in L. This one-sided mismatch can be compared with the ideal condition of two-sided match to accomplish the objective of our investigation. For example, we tried to generate sentence:

"二零零一年一月二日, 下午三點三十三分十一秒"

whose concatenated syllable segments should be

/ji-ling-ling-jat1 (1)nin4(1) jat1 (1)jyut6(3) baat3 (3)jat6, haa-ng5(1) saam1 (1)dim2(3) sei3-sap-gau2 (2)fan1(6) sap6-ng5 (5)miu5./
(key units corresponding to "年", "月", "日", "點", "分" and "秒" are underlined)

in ideal case where the tonal context always matches on both sides. If we introduce a mismatch on the left tonal context but maintain matches on the right, a possible concatenated syllable segment is:

/ji-ling-ling-jat1 (2)nin4(1) jat1 (3)jyut6(3) baat3 (4)jat6, haa-ng5(1) saam1 (5)dim2(3) sei3-sap-gau2 (6)fan1(6) sap6-ng5 (1)miu5./

We have 36 generated sentences with mismatched left tonal contexts and matching right tonal contexts. If we introduce a mismatch on the right tonal context but maintain matches on the left, a possible concatenated syllable segment is:

/ji-ling-ling-jat1 (1)nin4(2) jat1 (1)jyut6(4) baat3 (3)jat6, haa-ng5(6) saam1 (1)dim2(1) sei3-sap-gau2 (2)fan1(3) sap6-ng5 (5)miu5./

We have 36 generated sentences with mismatched right tonal contexts and matching left tonal contexts.

- Waveform evaluation – The comparison between a one-sided mismatch with a two-sided match is evaluated by an informal listening test with

eight subjects. Each subject listens 36 triplets, each triplet is ordered as:

1. concatenated syllable waveforms with matching tonal contexts on both sides,
2. concatenated syllable waveforms with matching left tonal context and mismatched right tonal context; and
3. concatenated syllable waveforms with matching right tonal context and mismatched left tonal context.

The subjects are told that waveform (1) offers the ideal generation quality, and are asked to compare (2) with (3) in achieving the quality that approaches (1). 36 waveform triplets for eight listeners gave us 288 ratings. Over 90% of the ratings claim that (2) is better than (3). Hence we conclude that the left tonal context has greater influence than right tonal context for speech generation by syllable concatenation.

5.2.2 Tonal Contexts in the Numeric Subgrammar

In other to corroborate our findings from the DATE-TIME subgrammar, we investigate the relative importance between left and right tonal contexts with the NUMERIC subgrammar as well. This subgrammar generated three-digit numeric string outputs, e.g. "一二三" (i.e. one two three). We chose this subgrammar for two reasons. The first one is the full coverage of the

six Cantonese tones in the Chinese numeric characters zero(零) to nine(九). The pronunciation of Chinese numeric characters are shown in Table 5.2.

Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Tone 6
一/jat1/ 七/cat1/	九/gau2/	三/saam3/ 四/sei3/ 八/baat3/	零/ling4/	五/ng5/	二/ji6/ 六/luk6/

Table 5.2: Chinese numeric syllables (0-9) categorized with the six tones.

The second reason is that we already have an existing wavebank for Chinese numeric syllables. The wavebank has a certain coverage of tone variants for each syllable. The syllables units are concatenated to generate three-digit numeric strings. Short strings are favorable in this investigation because they help listeners to concentrate on the quality of the target tonal syllables (in the middle of the triplet). To compare the importance of the left and right tonal contexts, two categories of waveforms pairs are generated, with 8 pairs in each category.

- In the first category, we generate two waveforms for the same digit string. The syllables for the first and third digits are the same across the waveform pair. For the middle syllable, we ensure that it has matching left context for one of the waveforms (denoted as MATCHED_LEFT), and mismatched left context for the other (denoted as MISMATCHED_LEFT). Consider the example of an given digit string "一二三" (i.e. one two three), which is pronounced as /jat1 ji6 saam1/. The first waveform concatenates /jat1 (1)ji6(1) saam1/ (note that the left tonal context of

/(**1**)ji6(1)/ matches the tone of its left neighbour /jat**1**/) The second waveform concentrates /jat1 (2)ji6(1) saam1/ (note that the left tonal context of /(**2**)ji6(1)/ does not match the tone of its left neighbour /jat**1**/).

- In the second category, the waveform pairs have the same format as the first, except that we are varying the right tonal context. We denote this pair as MATCHED_RIGHT and MISMATCHED_RIGHT.

A listening test was setup as a within-group experiment. 72 university students aged between 20 to 25 were invited as our listeners. Several precautions were taken to ensure a fair and unbiased environment for the listening test. For example, the order of two waveforms in each waveform pairs and the order of the waveform pairs in the whole listening test are randomized to eliminate learning effects. Each pair of waveforms were played three times before the listeners were asked to write down their judgments.

For each waveform pair, each listener is asked input one of the following:

- MATCHED_LEFT > MISMATCHED_LEFT refers to the former sounding better than the latter;
- MATCHED_LEFT = MISMATCHED_LEFT refers to the former sounding equally well as the latter;
- MATCHED_LEFT < MISMATCHED_LEFT refers to the former sounding worse than the latter;

Results from our 72 listeners judging 8 waveform pairs (of matched and mismatched left tonal contexts) gave 576 ratings, distributed as as follows:

Judgment	No. of ratings
MATCHED_LEFT=MISMATCHED_LEFT	93
MATCHED_LEFT>MISMATCHED_LEFT	377
MATCHED_LEFT<MISMATCHED_LEFT	106

We conducted statistical tests to analyze these results. A two-tailed test established with significance level (α=0.01) that there is perceivable difference between MATCHED_LEFT and MISMATCHED_LEFT. See Appendix D for details.

Another statistical test is a one-tailed test that focused only on the subset of waveforms with perceivable differences in their ratings. This test established with significance level (α=0.01) that listeners prefer MATCHED_LEFT over MISMATCHED_LEFT. See Appendix D for details.

Results from our 72 listeners judging 8 waveform pairs (of matched and mismatched right tonal contexts) gave 576 ratings, distributed as as follows:

Judgment	No. of ratings
MATCHED_RIGHT=MISMATCHED_RIGHT	259
MATCHED_RIGHT>MISMATCHED_RIGHT	177
MATCHED_RIGHT<MISMATCHED_RIGHT	140

We conducted statistical tests to analyze these results. A two-tailed test established with significance level (α=0.01) that there is *no* perceivable dif-

ference between MATCHED_RIGHT and MISMATCHED_RIGHT. See Appendix E for details.

Another statistical test is a one-tailed test that focused only on the subset of waveforms with perceivable differences in their ratings. This test established with significance level (α=0.01) that there is no significant consistency in listeners claiming preference to MATCHED_RIGHT over MISMATCHED_RIGHT or vice versa. See Appendix E for details.

5.2.3 Conclusion regarding the Relative Importance of Left versus Right Tonal Contexts

This section presents our study comparing the influence of the left tonal contexts versus the right in modifying the tone shape of a syllable. Our investigation with the scopes of the DATE-TIME and NUMERIC subgrammars corroborate one another, and show the left tonal context is much more important than the right one.

5.3 Selection Scheme for Tonal Variants

Our approach to Chinese speech generation based on syllable concatenation should ideally take both the coarticulatory and tonal contexts into account for unit selection. Even though we established above that perhaps only the left tonal context needs to be considered, we still need many tonal variants in order to guarantee that a syllable waveform with matching coarticulatory

and left tonal context can always be found during concatenation. It may not be realistic to expect that the wavebank can contain all such variants, hence we believe it is necessary to develop a "backoff scheme" for unit selection in substituting a missing tonal variant with reasonably good alternatives.

Following the previous arguments, if we need to select a syllable for concatenation, we should choose the tonal variant with matching left tonal context and right tonal context. If this tonal variant cannot be found, then we will try enforce only a matching left tonal context and disregard the right tonal context. However, if even such a condition cannot be met, we will revert to our "backoff scheme" that is developed with four prioritized principles in mind. The principles are derived mainly based on the observations of tone distortion examples from the previous experiments. The four prioritized principles are listed below:

1. Assume that the target tonal variant for concatenation is the syllable SYLT with tone T, and with desired left tonal context L_D (i.e. we can denote this syllable unit as $(L_D)\text{SYL}T$). The difference in tone height observed in this target tonal variant is $d = T - L_D$. If we denote the tonal variant substitute for concatenation as (L_S)SYLT, where L_S is the substituted left tonal context. The difference in tone height observed in this tonal variant substitute is $d' = T - L_S$. Then the substitute is chosen such that d' and d have the same sign (positive or negative). This maintains the slope in the tone trajectory as we move from the preceding syllable to the current one. This is illustrated in Figure 5.6.

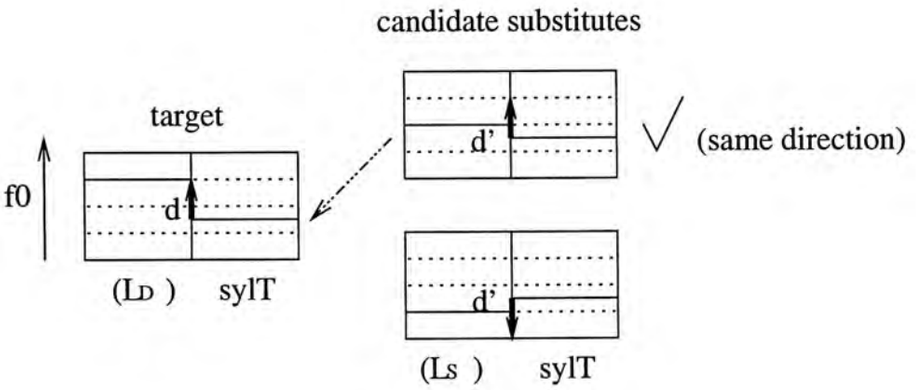


Figure 5.6: To find a tone variant substitute $(L_S)\text{SYL}T$ for $(L_D)\text{SYL}T$, we comparing the sign of d $(T - L_D)$ and d' $(T - L_S)$, and select $(L_S)\text{SYL}T$ such that d and d' has the same sign.

2. Given the above condition is satisfied, then we favor tonal variant substitutes whose left tonal context L_S has the same tone shape as L_D. Recall that there are two tone shapes – *rising* e.g. for tones 2 and 5; and *level* for tones 1, 3, 4 and 6. This is illustrated in Figure 5.7.

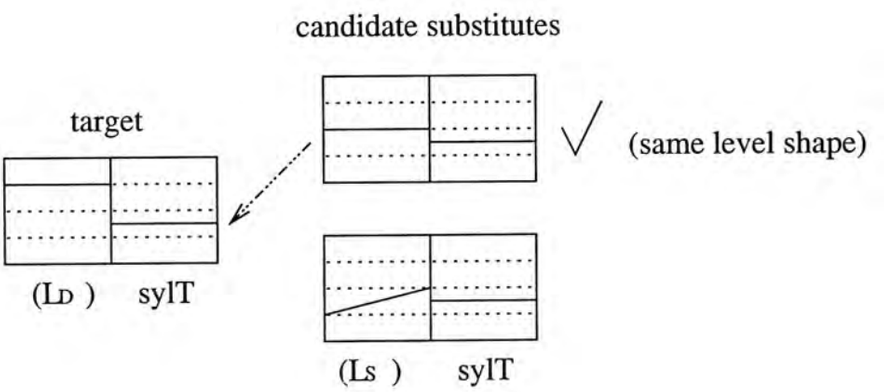


Figure 5.7: To find a tone variant substitute $(L_S)\text{SYL}T$ for $(L_D)\text{SYL}T$, we comparing the tone shapes of L_D and L_S, and we favor $(L_S)\text{SYL}T$ whose L_S has the same tone shape as L_D.

3. Given that (1) and (2) have been satisfied, then if d is negative, we will

favor tonal variant substitutes with d' values which are *less negative* over those with d' values which are *more negative.* Thereafter we aim to minimize the magnitude difference between d' and d. By the same token, if d is positive, then we favor tonal variants substitutes with d' values which are *less positive* over those with d' values which are *more positive.* Thereafter we aim to minimize the magnitude difference between d' and d. This principle avoids large transitional movements in the tone trajectory going from the preceding syllable to the current syllable. This is illustrated in Figure 5.8.

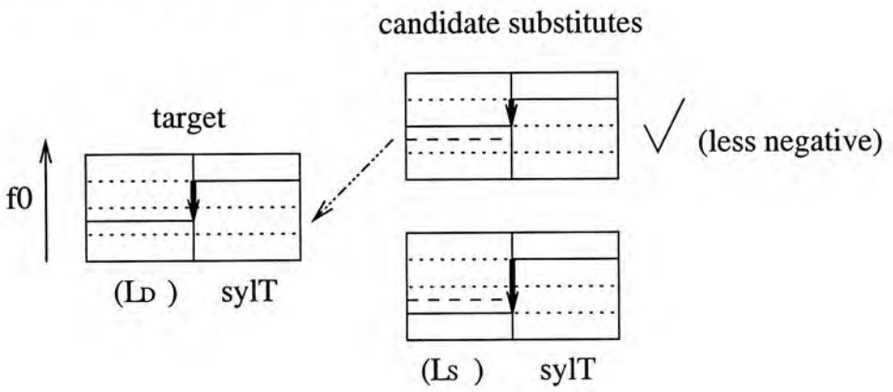


Figure 5.8: To find a tone variant substitute $(L_S)\text{SYL}\,T$ for $(L_D)\text{SYL}\,T$, we comparing the magnitude of d $(T - L_D)$ and d' $(T - L_S)$. If d is positive, we favor $(L_S)\text{SYL}\,T$ with a less positive d'. If d is negative, we favor $(L_S)\text{SYL}\,T$ with less negative d'. Thereafter we aim to minimize the magnitude difference between d' and d.

4. We try to avoid tone variant substitutes whose L_S is tone 2. This is because tone 2 has one of the most dynamic tone shapes and often leads to overshooting and undershooting in tone trajectories. This is illustrated in Figure 5.9.

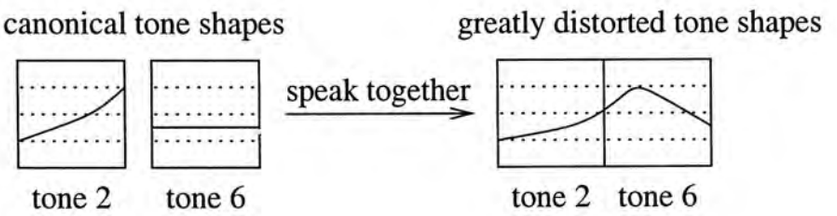


Figure 5.9: Overshooting and undershooting in tone trajectories when L_S is tone 2.

5.3.1 Listening Test for our Tone Backoff Scheme

We designed a listening test to assess the validity of our "tone backoff scheme". This is based on Chinese speech generation of digit triplets, similar to the listening test in the previous section. The digit triplet calls for a tonal variant for the middle syllable with tone T and desired left context L_D. However, we try to substitute another variant with tone T and left context L_S according to our backoff scheme. We have used 15 digit triplets, covering five tones for T and five tones for L_D, but not all combinations are included due to limitations in our wavebank. For a given T and L_D, our backoff scheme provides a rank order of up to five alternatives for L_S. Of these five alternatives, we include three in our generated waveforms. Hence, for each digit string, we generate a group of four waveforms:

For example, given a digit string "九一三" (i.e. nine one three) pronounced as /gau2 jat1 saam1/:

- the first generated waveform is /gau2(1) (2)jat1(1) (1)saam1/ (ideal case with matching tonal variants, denoted as REF)

- the second generated waveform is /gau2(1) (5)jat1(1) (1)saam1/ (replaced the $L_D = 2$ with $L_S = 5$) (denoted as SUBSTITUTE1)

- the third generated waveform is /gau2(1) (4)jat1(1) (1)saam1/ (replaced the $L_D = 2$ with $L_S = 4$) (denoted as SUBSTITUTE2)

- the fourth generated waveform is /gau2(1) (1)jat1(1) (1)saam1/ (replaced the $L_D = 2$ with $L_S = 1$) (denoted as SUBSTITUTE3)

We played the REF first, and then the other three waveform in randomized order. Listeners were asked to rank SUBSTITUTE1, SUBSTITUTE2, SUBSTITUTE3 in the descending order of naturalness (please see Appendix F for the listening test questionnaire). We then compare the listener's ranking with that suggested by the backoff scheme. This is done one pair at a time. For example, if the listener's ranking is: SUBSTITUTE1 > SUBSTITUTE2 = SUBSTITUTE3 then the pairs are (1>2), (1>3) and (2=3).

This compares with our backoff scheme and as an example, if the scheme suggests the ranking: SUBSTITUTE1 > SUBSTITUTE3 > SUBSTITUTE2 and then the pairs are (1>2), (1>3) and (3>2).

So comparison between the two rankings by means of the generated pairs will show two pairs in agreement ($N_{agree} = 2$) out of three pairs in total and one pair with no perceivable difference. We also denote the pair in disagreement as ($N_{disagree} = 0$).

Recall that we have 15 digit triplets, and their 45 waveforms are all rated by the 55 listeners as described above, giving 2475 pairs of ratings. Results

Judgment	No. of ratings
No perceivable differences	300
N_{agree}	1650
$N_{disagree}$	525

Table 5.3: Results from the listening test for the validity of the "backoff scheme".

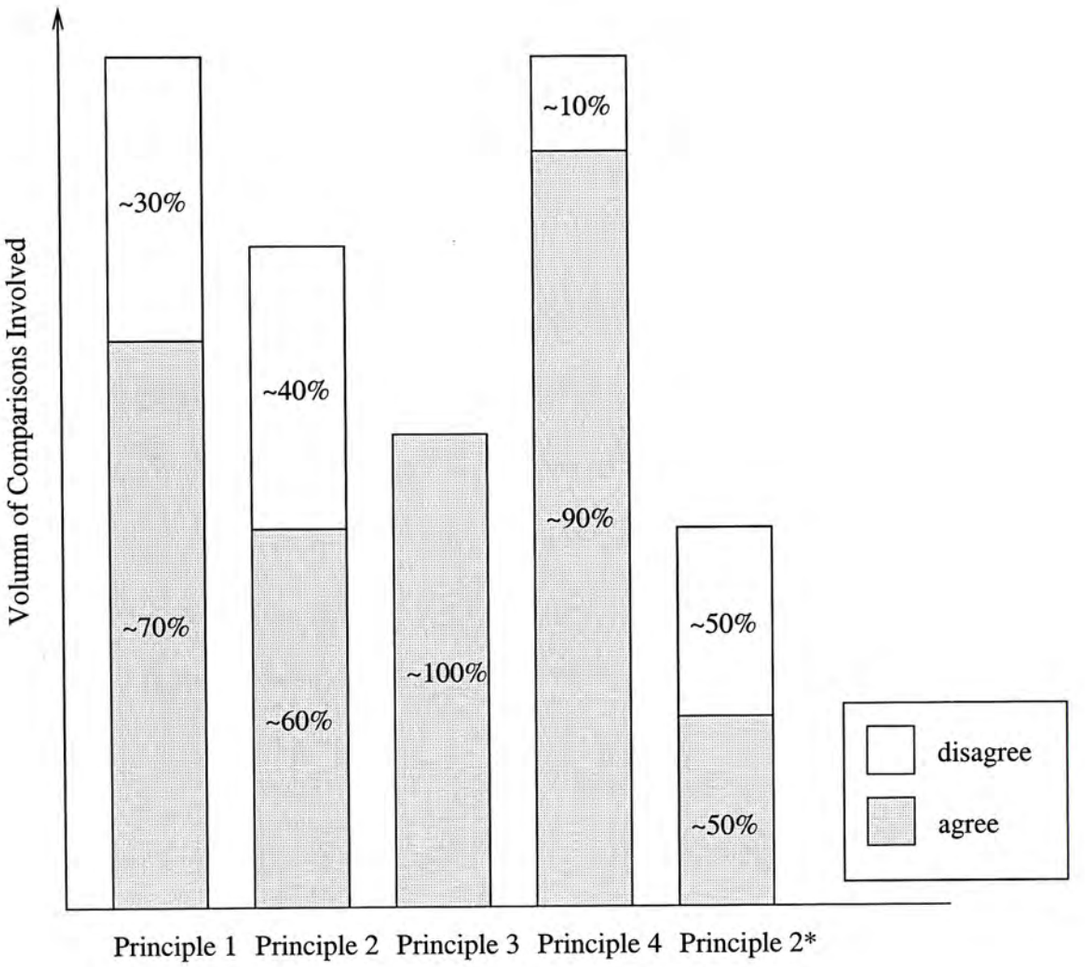
are tabulated in Table 5.3.

A statistical test with significance level $\alpha = 0.01$ established that the probability of agreement between the listener's perception and our backoff scheme is greater than 70%. Details are given in Appendix G.

5.3.2 Error Analysis

Error analysis is performed on the experimental results to check the validity of each principle (see Section 5.3) applied in the backoff scheme. To evaluate the validity of each principle, we consider only the comparisons which are directly relevant for the principle. This gives around 500 comparisons for each principle. Among these comparisons, we count the number of comparisons that listeners' claims *agree* with the principle, as well as the number of comparisons that listener's claims *disagree* with the principle.

The principle is considered to be valid if the number of *agree* dominates over the number of *disagree*. As shown in Figure 5.10, we see that approximately 70% of the comparisons support the Principle 1 (i.e. that we need to maintain the slope in the tone). About 60% of the listening comparisons



Principle 1: maintain the slope
Principle 2: maintain the tone shape
Principle 3: avoid large transitionaL movements
Principle 4: avoid substituting with tone 2 as left context
Principle 3*: apply principle 3 when all substitutes do not follow principle 1

Figure 5.10: Proportion of comparisons that agree or disagree with the principles.

support Principle 2 (i.e. we should maintain the tone shape in finding a substitute for the missing tonal variant). 100% of the listening comparisons support Principle 3 (i.e. we should avoid large transitional movements in the tone trajectory going from the preceding syllable to the current syllable). About 90% of the listening comparisons support Principle 4 (i.e. we should avoid substituting with tone 2 as left context). Our backoff scheme also specifies that if we cannot find a substitute that follows Principle 1, then we should move on directly to Principle 3. This part of the scheme is only supported by about 50% of the listening comparisons. Statistical tests with significance level $\alpha = 0.01$ established that the four principles are valid.

5.4 Chapter Summary

In this chapter, we described our investigation on the effect of tonal context towards the naturalness of generated speech. Several experiments are conducted regarding the left and right tonal context. Our investigation suggests that the left tonal context has greater influence than the right tonal context over the tone trajectory of the current syllable for concatenative synthesis. It is often difficult to create a wavebank that can always provide a syllable with matching tonal and coarticulatory context for concatenation. Hence we developed a backoff scheme to find tonal variant substitutes when the desired variant is missing. A listening test suggests that our backoff scheme agrees with human perception 70% of the time.

Chapter 6

Summary and Future Work

This thesis describes an approach for domain-optimized speech generation in Chinese, which is to generate speech with optimized naturalness and intelligibility within the scope of an application domain. Our research goals are (i) to develop a response generation framework for Chinese, which is able to produce highly natural synthesized speech and can be applied to different Chinese dialects (Mandarin and Cantonese); (ii) to enhance the framework for the portability across application domains and the scalability to more complex domains (e.g. stocks); and (iii) to improve the speech quality specially for tonal distortion so that the framework can produce highly nature speech output in complex domains. In other to achieve these goals, we have completed three major tasks. We developed a framework of concept-to-speech response generation for information delivery with the approach of corpus-based syllable concatenation. We started by testing the feasibility of

the framework in the small application domain, and ensured the framework can produce highly natural and intelligible speech output by carefully design the recording corpus and capturing coarticulatory context with distinctive features. We proved by a listening test that the speech output generated with the framework is highly natural. As a next step, we ported the framework to a more complex domain for investigating portability and scalability. The framework is generalized for response synthesis across domains with enhancements in the representation of input semantics and response grammars, the reusability of grammars, the algorithm for recording corpus development and smoothening in the energy level of the speech output. We continued the work by investigating the influence of tonal context. The investigation is motivated by the observation that prosodic distortions of pitch levels are perceived in the synthesized speech when the framework is applied on the more complex domain. We investigated the influence of tonal context in several hierarchies. Investigation on the relative importance between the left and right tonal contexts is first carried out. From the experimental results we observed that the distortion caused by left tonal context is more significant. Based on the distortions, we developed a set of rules to predict human perception on naturalness in terms of the unit selection on tonal context, and we developed a unit selection scheme for tonal context based on the rules. Analysis showed that the unit selection scheme successfully predicted human perception with an accuracy of 70%.

6.1 Contributions

In this thesis, the following contributions are made to the research area of Chinese speech generation.

1. A single framework is developed for domain-optimized speech generation for both Mandarin and Cantonese. In our approach, coarticulatory context is considered with the use of linguistic-motivated distinctive features to ensure highly natural and intelligible speech output. A generate-and-filter algorithm is developed to automatically produce compact set of recording prompts which fully cover the vocabulary needed within the scope of domain [39].

2. The framework is enhanced to be portable and scalable across application domains. XML technology is applied to achieve domain portability and scalability. An alternate approach, namely the tree-based filtering algorithm, is developed for recording corpus development. This enhanced approach eases the bottleneck of the generate-and-filter algorithm in large application domains. The technology of energy normalization is applied to smoothen the energy level of the generated speech, hence increase naturalness [4, 5].

3. A empirical study on the relative importance between left and right tonal context towards the naturalness of speech output by syllable based concatenation.

4. A set of rules for human perception on tonal context is developed and a unit selection scheme regarding tonal context is developed accordingly to predict human perception towards naturalness.

6.2 Future Directions

This thesis demonstrated the feasibility, scalability and portability of the response generation framework for Chinese, also the feasibility of the approach for prosody modeling on pitch with investigations on tonal context. There are number of possible directions for extension of this work. In this section, we list some possible directions for future work.

One direction is to integrate the unit selection scheme on tonal context to the framework as an enhancement. Experimental results show that our unit selection scheme can predict human perception at 70% accuracy, however, it has not been applied in a real system. Since we have developed a number of wavebanks for different applications domains, the unit selection scheme on tonal context can augment with the existing unit selection algorithm to account for both tonal and coarticulatory contexts during concantenation. In this way, in additional to ensure units selected with criteria of matching coarticulatory context, the units with most preferred tonal context are selected within the existing wavebanks the possible candidates of corresponding acoustic units for concatenation. The enhancement should be able to minimized the distortion caused by tonal context. An alternate method for

the integration is to consider the unit selection scheme for tonal context in recording corpus development, by ensuring the variations of most preferred tonal context of the vocabulary are covered in the recording prompts.

A third direction of work is to further investigate the effects of tonal context to increase the accurately of human perception prediction. Our unit selection scheme succeeded with fairly high accuracy, but there are still room for improvement. In this thesis we considered the tonal context as the tone of immediately preceding and immediately following syllables. Investigations on tonal context can be extend to consider a larger number of neighboring syllables and also the positioning of the syllable in the whole sentence.

Another direction regarding to the flexibility of the framework is handling out-of-vocabulary (OOV) words. This can be achieved by using a very large corpus of speech for a wavebank, which covered all mono-syllables units and at the same time cover as many variations of coarticulatory context and tonal context as possible. The wavebank from the very large corpus can be created as an additional wavebank for OOV handling. An enhancement on OOV detection can be made on the response generation framework. When an OOV is detected, mono-syllable units will be selected from the additional wavebank to ensure the a complete response can be generated. The technique of OOV handling provide the flexibility for our response generation framework to handle more open domains, such as financial news.

Appendix A

Listening Test Questionnaire for FOREX Response Generation

Listening Test Questionnaire

Instructions
Please go to the URL:
http://www.se.cuhk.edu.hk/~tyfung/abc/ab.html

There are 10 icons there. Click each icon to listen to each of the sentences.
All the sentences together will cover all the foreign currencies worldwide.
Please rate each sentence according to the intelligibility and naturalness below.

Rating Scale of 1 to 6
1 (barely intelligible) 2 3 4 5 6 (very intelligible)
1 (barely natural) 2 3 4 5 6 (very natural)

Sentence 1	Intelligibility_____(left)	Naturalness_____(left)
	Intelligibility_____(right)	Naturalness_____(right)
	Comments_____	
Sentence 2	Intelligibility_____(left)	Naturalness_____(left)
	Intelligibility_____(right)	Naturalness_____(right)
	Comments_____	
Sentence 3	Intelligibility_____(left)	Naturalness_____(left)
	Intelligibility_____(right)	Naturalness_____(right)
	Comments_____	
Sentence 4	Intelligibility_____(left)	Naturalness_____(left)
	Intelligibility_____(right)	Naturalness_____(right)
	Comments_____	
Sentence 5	Intelligibility_____(left)	Naturalness_____(left)
	Intelligibility_____(right)	Naturalness_____(right)
	Comments_____	
Sentence 6	Intelligibility_____(left)	Naturalness_____(left)
	Intelligibility_____(right)	Naturalness_____(right)
	Comments_____	
Sentence 7	Intelligibility_____(left)	Naturalness_____(left)
	Intelligibility_____(right)	Naturalness_____(right)
	Comments_____	
Sentence 8	Intelligibility_____(left)	Naturalness_____(left)
	Intelligibility_____(right)	Naturalness_____(right)
	Comments_____	
Sentence 9	Intelligibility_____(left)	Naturalness_____(left)
	Intelligibility_____(right)	Naturalness_____(right)
	Comments_____	
Sentence 10	Intelligibility_____(left)	Naturalness_____(left)
	Intelligibility_____(right)	Naturalness_____(right)
	Comments_____	

Figure A.1: The listening test questionnaire to evaluate the speech output of FOREX response generation.

Appendix B

Major Response Types For ISIS

Response Types in Cantonese

類別一) 買賣證券
i) 買入證券
請確認你o既指示：買入0005匯豐控股四百股，每股一百蚊。你而家係咪要執行呢個指示呢？
ii)沽出證券
請確認你o既指示：賣出0005匯豐控股四百股，每股一百蚊。你而家係咪要執行呢個指示呢？

類別二) 查詢買賣指示處理情況
i) 查詢買入指出
你於二零零零年八月廿四日十一點三十分買入0005匯豐控股四百股，每股 一百蚊。
ii)查詢沽出指出
你於二零零零年八月廿四日十一點三十分賣出0005匯豐控股四百股，每股 一百蚊。
iii)查詢其他服務
同上，總結過去三個交易日買入賣出情況。

類別三) 更改買賣指示
請確認你o既指示：更改賣出/取消買入0005匯豐控股四百股，每股一百蚊。
新指示係賣出/新指示係買入0005匯豐控股三百股，每股一百蚊。你而家係咪要執行呢個指示呢？

類別四) 取消買賣指示
請確認你o既指示：
取消賣出/取消買入0005匯豐控股四百股，每股一百蚊。
取消賣出/取消買入0005匯豐控股一千股，每股九十九蚊。
你而家係咪要執行呢個指示呢？

類別五) 查詢即時股票資料
> 0005匯豐控股成交價/昨日收市價/今早開市價/今日最高價/今日最低價係 一百蚊。
> 0005匯豐控股總成交量係四百萬股。
> 0005匯豐控股每股上升/下跌一蚊。
> 0005匯豐控股無升跌。

類別六) 查詢即時指數及最新市場消息
> 現在恒生指數/上海A及B股指數/深圳A及B股指數係一萬二千點。
> 0005匯豐控股每股上升/下跌一蚊。
> 0005匯豐控股每股派息零點四蚊。
> 0005匯豐控股 股數發行量為五百萬股。
> 0005匯豐控股上一次股息發行日期係二零零零年七月七日。
> 十大活躍股票/十大成交量股票/十大升幅股票/十大跌幅股票係0005匯豐控股每股五十蚊，上升/下跌一蚊、0011恒生銀行每股六十蚊，上升/下跌零點五蚊…… …

類別七) 查詢戶口中所有或個別證券存貨
> 你持有0005匯豐控股一千股，每股盈利/虧損五蚊。
> 你持有0011恒生銀行一千股，無任何盈利或虧損。

類別八) 查詢公司的最新新聞標題
> 對唔住，0005匯豐控股無新聞。
> 0005匯豐控股有一段新聞請睇。 (Display news on screen)

類別九) 查詢股票的最新走勢圖
0005匯豐控股最新走勢圖，請睇。 (Display graph on screen)

Figure B.1: Response types of ISIS in Cantonese.

Response Types in Mandarin

類別一) 買賣證券
i) 買入證券
請確認你的指示：買入 0005 匯豐控股四百股，每股一百元。你現在要執行這 個指示嗎？
ii)沽出證券
請確認你的指示：賣出 0005 匯豐控股四百股，每股一百元。你現在要執行這個指示嗎？

類別二) 查詢買賣指示處理情況
i) 查詢買入指出
你於一九九九年八月廿四日十一時三十分買入 0005 匯豐控股四百股，每股一百元。
ii)查詢沽出指出
你於一九九九年八月廿四日十一時三十分賣出 0005 匯豐控股四百股，每股一百元。
iii)查詢其他服務
同上，總結過去三個交易日買入賣出情況。

類別三) 更改買賣指示
請確認你的指示：更改賣出/更改買入 0005 匯豐控股四百股，每股一百元。
新指示為賣出/新指示為買入 0005 匯豐控股三百股，每股一百元。你現在要執行這個指示嗎？

類別四) 取消買賣指示
請確認你的指示：
取消賣出/取消買入 0005 匯豐控股四百股，每股一百元。
取消賣出/取消買入 0005 匯豐控股一千股，每股九十九元。
你現在要執行這個指示嗎？

類別五) 查詢即時股票資料
> 0005 匯豐控股成交價是/昨日收市價是/今早開市價是/今天最高價是/今天最低價是一百元。
> 0005 匯豐控股總成交量是四百萬股。
> 0005 匯豐控股每股上升/下跌一元。
> 0005 匯豐控股無升跌。

類別六) 查詢即時指數及最新市場消息
> 現在恒生指數/上海 A 及 B 股指數/深圳 A 及 B 股指數是一萬二千點。
> 0005 匯豐控股每股上升/下跌一元。
> 0005 匯豐控股每股派息零點四元。
> 0005 匯豐控股股數發行量為五百萬股。
> 0005 匯豐控股上一次股息發行日期為一九九九年七月七日。
> 十大活躍股票/十大成交量股票/十大升幅股票/十大跌幅股票是 0005 匯豐控股，每股五十元，上升/下跌一元、0011 恒生銀行，每股六十元，上升/下跌零點五元 … …

類別七) 查詢戶口中所有或個別證券存貨
> 你持有 0005 匯豐控股一千股，每股盈利/虧損五元。
> 你持有 0011 恒生銀行一千股，無任何盈利或虧損。

類別八) 查詢公司的最新新聞標題
> 對不起，0005 匯豐控股無新聞。
> 0005 匯豐控股有一段新聞，請看。 (Display news on screen)

類別九) 查詢股票的最新走勢圖
0005 匯豐控股的最新走勢圖，請看。 (Display graph on screen)

Figure B.2: Response types of ISIS in Mandarin.

Appendix C

Recording Corpus for Tone Investigation in Date-time Subgrammar

二零零一年一月二日，下午三點三十三分十一秒
二零零三年九月六日，下午一點三十九分十二秒
二零零七年四月十日，上午九點三十八分十三秒
二零零一年五月二日，下午四點三十五分十四秒
二零零三年二月六日，下午五點三十分十五秒
二零零七年三月十日，下午六點三十七分十六秒
二零零九年七月一日，上午十點三十一分九秒
二零零九年九月三日，下午一點三十九分九秒
二零零九年八月七日，上午七點三十四分九秒
二零零九年五月一日，上午九點十五分九秒
二零零九年六月三日，上午八點十二分九秒
二零零九年一月七日，下午兩點十三分九秒
二零零四年三月九日，下午五點四十七分三十七秒
二零零八年九月九日，上午六點四十九分三十八秒
二零零四年四月九日，上午七點四十八分三十九秒
二零零八年五月九日，上午三點四十五分三十秒
二零零四年十月九日，上午九點四十六分三十一秒
二零零八年七月九日，下午四點四十一分三十二秒
二零零零年一月十四日，上午八點零三分四十一秒
二零零零年九月二十九日，下午兩點零九分四十二秒
二零零零年八月三十一日，上午十點零四分四十三秒
二零零零年五月十七日，下午三點零五分四十四秒
二零零零年二月二十五日，下午一點零六分四十五秒
二零零零年三月三十日，上午九點零一分四十六秒
二零零五年七月四日，上午九點五十一分零一秒
二零零五年九月八日，下午四點五十九分零二秒
二零零五年四月四日，下午五點五十八分零三秒
二零零五年五月八日，下午六點五十五分零四秒
二零零五年六月四日，下午一點五十二分零五秒
二零零五年一月八日，上午七點五十三分零六秒
二零零二年三月五日，上午七點二十七分五十七秒
二零零六年九月五日，下午三點二十九分五十八秒
二零零二年八月五日，上午九點二十四分五十九秒
二零零六年五月五日，上午八點二十五分五十秒
二零零二年十月五日，下午兩點二十六分五十一秒
二零零六年七月五日，上午十點二十一分五十二秒

Table C.1: Recording prompts for value units in date-time subgrammar.

我要讀 一 年 一 同 一 月 一 俾你聽
我要讀 一 年 險 同 一 月 險 俾你聽
我要讀 一 年 四 同 一 月 四 俾你聽
我要讀 一 年 紅 同 一 月 紅 俾你聽
我要讀 一 年 社 同 一 月 社 俾你聽
我要讀 一 年 十 同 一 月 十 俾你聽
我要讀 九 年 一 同 九 月 一 俾你聽
我要讀 九 年 險 同 九 月 險 俾你聽
我要讀 九 年 四 同 九 月 四 俾你聽
我要讀 九 年 紅 同 九 月 紅 俾你聽
我要讀 九 年 社 同 九 月 社 俾你聽
我要讀 九 年 十 同 九 月 十 俾你聽
我要讀 四 年 一 同 四 月 一 俾你聽
我要讀 四 年 險 同 四 月 險 俾你聽
我要讀 四 年 四 同 四 月 四 俾你聽
我要讀 四 年 紅 同 四 月 紅 俾你聽
我要讀 四 年 社 同 四 月 社 俾你聽
我要讀 四 年 十 同 四 月 十 俾你聽
我要讀 團 年 一 同 團 月 一 俾你聽
我要讀 團 年 險 同 團 月 險 俾你聽
我要讀 團 年 四 同 團 月 四 俾你聽
我要讀 團 年 紅 同 團 月 紅 俾你聽
我要讀 團 年 社 同 團 月 社 俾你聽
我要讀 團 年 十 同 團 月 十 俾你聽
我要讀 五 年 一 同 五 月 一 俾你聽
我要讀 五 年 險 同 五 月 險 俾你聽
我要讀 五 年 四 同 五 月 四 俾你聽
我要讀 五 年 紅 同 五 月 紅 俾你聽
我要讀 五 年 社 同 五 月 社 俾你聽
我要讀 五 年 十 同 五 月 十 俾你聽
我要讀 二 年 一 同 二 月 一 俾你聽
我要讀 二 年 險 同 二 月 險 俾你聽
我要讀 二 年 四 同 二 月 四 俾你聽
我要讀 二 年 紅 同 二 月 紅 俾你聽
我要讀 二 年 社 同 二 月 社 俾你聽
我要讀 二 年 十 同 二 月 十 俾你聽

Table C.2: Recording prompts for key units "年" and "月". Exhaustive tonal context is created for "年" and "月" by their neighboring tonal syllables. The phrases "我要讀" and "俾你聽" is added to minimize sentential declination effects and prepausal lengthenings.

我要讀 一 點 一 同 一 分 一 俾你聽
我要讀 一 點 險 同 一 分 險 俾你聽
我要讀 一 點 四 同 一 分 四 俾你聽
我要讀 一 點 紅 同 一 分 紅 俾你聽
我要讀 一 點 社 同 一 分 社 俾你聽
我要讀 一 點 十 同 一 分 十 俾你聽
我要讀 九 點 一 同 九 分 一 俾你聽
我要讀 九 點 險 同 九 分 險 俾你聽
我要讀 九 點 四 同 九 分 四 俾你聽
我要讀 九 點 紅 同 九 分 紅 俾你聽
我要讀 九 點 社 同 九 分 社 俾你聽
我要讀 九 點 十 同 九 分 十 俾你聽
我要讀 四 點 一 同 四 分 一 俾你聽
我要讀 四 點 險 同 四 分 險 俾你聽
我要讀 四 點 四 同 四 分 四 俾你聽
我要讀 四 點 紅 同 四 分 紅 俾你聽
我要讀 四 點 社 同 四 分 社 俾你聽
我要讀 四 點 十 同 四 分 十 俾你聽
我要讀 團 點 一 同 團 分 一 俾你聽
我要讀 團 點 險 同 團 分 險 俾你聽
我要讀 團 點 四 同 團 分 四 俾你聽
我要讀 團 點 紅 同 團 分 紅 俾你聽
我要讀 團 點 社 同 團 分 社 俾你聽
我要讀 團 點 十 同 團 分 十 俾你聽
我要讀 五 點 一 同 五 分 一 俾你聽
我要讀 五 點 險 同 五 分 險 俾你聽
我要讀 五 點 四 同 五 分 四 俾你聽
我要讀 五 點 紅 同 五 分 紅 俾你聽
我要讀 五 點 社 同 五 分 社 俾你聽
我要讀 五 點 十 同 五 分 十 俾你聽
我要讀 二 點 一 同 二 分 一 俾你聽
我要讀 二 點 險 同 二 分 險 俾你聽
我要讀 二 點 四 同 二 分 四 俾你聽
我要讀 二 點 紅 同 二 分 紅 俾你聽
我要讀 二 點 社 同 二 分 社 俾你聽
我要讀 二 點 十 同 二 分 十 俾你聽

Table C.3: Recording prompts for key units "點" and "分". The format follows Table C.2.

Appendix D

Statistical Test for Left Tonal Context

The parameter of interest is the probability that MATCHED_LEFT = MISMATCHED_LEFT, p
$H_0 : p = 0.5$
$H_1 : p \neq 0.5$
$\alpha = 0.01$
Using normal distribution to approximate binomial distribution, the test statistic is
$z_0 = \frac{p - E[p]}{\sqrt{Var[p]}}$
Reject H_0 if $z_0 > 2.58$ or if $z_0 < -2.58$
Computations: observed $p = \frac{93}{576} = 0.161$, $E[p] = 0.5$, $Var[p] = \frac{p(1-p)}{n} = \frac{(0.5)(1-0.5)}{576} = 0.000434$
$z_0 = \frac{0.161 - 0.5}{\sqrt{0.000434}} = -16.27$
Conclusion: Since $z_0 = -16.27 < -2.58$ falls into the rejection region
we reject H_0 and conclude at the 0.01 level of significance that $p = 0.5$.

Figure D.1: Details of statistical test on perceivable difference for left tonal context.

The parameter of interest is the probability of listeners prefers MATCHED_LEFT over MISMATCHED_LEFT, q
$H_0 : q = 1/2$
$H_1 : q > 1/2$
$\alpha = 0.01$
Using normal distribution to approximate binomial distribution, the test statistic is
$z_0 = \frac{N_{A>B} - E[N_{A>B}]}{\sqrt{Var[N_{A>B}]}}$
Reject H_0 if $z_0 > 2.33$
Computations: observed $N_{A>B} = 377$, $E[N_{A>B}] = Kq = (483) \times (\frac{1}{2}) = 241.5$, $Var[N_{A>B}] = Kq(1-q) = (483) \times (\frac{1}{2}) \times (1-\frac{1}{2}) = 120.75$
$z_0 = \frac{377-241.5}{\sqrt{120.75}} = 12.33$
Conclusion: Since $z_0 = 12.33 < 2.33$ falls into the rejection region we reject H_0 and conclude at the 0.01 level of significance that $q = 1/2$.

Figure D.2: Details of statistical test on listeners' preference for left tonal context.

Appendix E

Statistical Test for Right Tonal Context

The parameter of interest is the probability that listeners pick MATCHED_RIGHT = MISMATCHED_RIGHT, p
$H_0 : p = 0.5$
$H_1 : p \neq 0.5$
$\alpha = 0.01$
Using normal distribution to approximate binomial distribution, the test statistic is
$z_0 = \frac{p - E[p]}{\sqrt{Var[p]}}$
Reject H_0 if $z_0 > 2.58$ or if $z_0 < -2.58$
Computations: observed $p = \frac{259}{576} = 0.450$, $E[p] = 0.5$, $Var[p] = \frac{p(1-p)}{n} = \frac{(0.5)(1-0.5)}{576} = 0.000434$
$z_0 = \frac{0.450 - 0.5}{\sqrt{0.000434}} = -2.4$
Conclusion: Since $z_0 = -2.4 > -2.58$ falls out of the rejection region
we CANNOT reject H_0. We conclude at the 0.01 level of significance that we cannot reject the claim of $p = 0.5$.

Figure E.1: Details of statistical test on perceivable difference for right tonal context.

The parameter of interest is the probability of listeners prefers MATCHED_RIGHT over MISMATCHED_RIGHT, q
$H_0 : q = 1/2$
$H_1 : q > 1/2$
$\alpha = 0.01$
Using normal distribution to approximate binomial distribution, the test statistic is
$z_0 = \frac{N_{A>B} - E[N_{A>B}]}{\sqrt{Var[N_{A>B}]}}$
Reject H_0 if $z_0 > 2.33$
Computations: observed $N_{A>B} = 177$, $E[N_{A>B}] = Kq = (317) \times (\frac{1}{2}) = 79.25$, $Var[N_{A>B}] = Kq(1-q) = (317) \times (\frac{1}{2}) \times (1-\frac{1}{2}) = 158.5$
$z_0 = \frac{177-158.5}{\sqrt{79.25}} = 2.08$
Conclusion: Since $z_0 = 2.08 < 2.33$ falls out of the rejection region we CANNOT reject H_0. We conclude at the 0.01 level of significance that we cannot reject the claim of $q = 1/2$.

Figure E.2: Details of statistical test on listeners' preference for right tonal context.

Appendix F

Listening Test Questionnaire for Backoff Unit Selection Scheme

Listening Test Questionnaire

Instructions
Please go to the URL:
http://www.se.cuhk.edu.hk/~tyfung/listening_test/
(access after Thurs 15 March, 2001)

There are 15 waveform groups there. Click each wave to listen to each of the synthesized sentences. Please click the prompt to listen to the standard waves and compare WITHIN each group (A, B, C) and rank their naturalness.

Ranking Scale of 1 to 3
1: Most natural (sound most like the standard)
3: Least natural (sound less like the standard)
e.g
Sentence pair 0 A_2__ B_1__ C_3__
Comments___B is better than A, and A is better than C_____

Sentence pair 1 A___ B___ C___
Comments _____

Sentence pair 2 A___ B___ C___
Comments _____

Sentence pair 3 A___ B___ C___
Comments _____

Sentence pair 4 A___ B___ C___
Comments _____

Sentence pair 5 A___ B___ C___
Comments _____

Sentence pair 6 A___ B___ C___
Comments _____

Sentence pair 7 A___ B___ C___
Comments _____

Sentence pair 8 A___ B___ C___
Comments _____

Sentence pair 9 A___ B___ C___
Comments _____

Figure F.1: The listening test questionnaire to evaluate the backoff unit selection scheme

Appendix G

Statistical Test for the Backoff Unit Selection Scheme

The parameter of interest is the probability that the scheme agree with listeners' claims, q
$H_0 : q = 0.7$
$H_1 : q > 0.7$
$\alpha = 0.01$
Using normal distribution to approximate binomial distribution, the test statistic is
$z_0 = \frac{N_{A>B} - E[N_{A>B}]}{\sqrt{Var[N_{A>B}]}}$
Reject H_0 if $z_0 > 2.33$
Computations: observed $N_{nodiff} = 300$, $N_{L \neq S} = 525$, $N_{L=S} = 1650$, $E[N_{L=S}] = Kq = (1650 + 525) \times (0.7) = 1522.5$, $Var[N_{L=S}] = Kq(1-q) = (1650 + 525) \times (0.7) \times (0.3) = 456.75$
$z_0 = \frac{1650 - 1522.5}{\sqrt{456.75}} = 5.97$
Conclusion: Since $z_0 = 5.97 > 2.33$ falls into the rejection region, we conclude at the 0.01 level of significance that $q > 0.7$.

Figure G.1: Details of statistical test on the backoff unit selection scheme.

Appendix H

Statistical Test for the Backoff Unit Selection Scheme

The parameter of interest is the probability that the scheme agree with listeners' claims, q
$H_0 : q = 0.7$
$H_1 : q > 0.7$
$\alpha = 0.01$
Using normal distribution to approximate binomial distribution, the test statistic is
$z_0 = \frac{N_{A>B} - E[N_{A>B}]}{\sqrt{Var[N_{A>B}]}}$
Reject H_0 if $z_0 > 2.33$
Computations: observed $N_{nodiff} = 300$, $N_{L \neq S} = 525$, $N_{L=S} = 1650$, $E[N_{L=S}] = Kq = (1650 + 525) \times (0.7) = 1522.5$, $Var[N_{L=S}] = Kq(1-q) = (1650 + 525) \times (0.7) \times (0.3) = 456.75$
$z_0 = \frac{1650 - 1522.5}{\sqrt{456.75}} = 5.97$
Conclusion: Since $z_0 = 5.97 > 2.33$ falls into the rejection region, we conclude at the 0.01 level of significance that $q > 0.7$.

Figure H.1: Details of statistical test on the backoff unit selection scheme.

Bibliography

[1] Yi, J. and Glass, J. "Natural-Sounding Speech Synthesis Using Variable-Length Units". In *Proceedings of International Conference on Spoken Language Processing*, 1998.

[2] Rabiner, L. R. and Schafer, R. W. *"Digital Processing of Speech Signals"*, pages 39–41. Prentice-Hall, 1978.

[3] Meng, H. M., Lee, S., and Wai, C. "CU FOREX: A bilingual Spoken Dialog System for the Foreign Exchange Domain". In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2000.

[4] Meng, H., Chan. S. F., Wong, Y. F., Fung. T. Y., Tsui, W. C., Lo, T. H., Chan, C. C., Chen, K. , Wang, L, Wu, T. Y., Li, X. L., Lee, T., Choi, W. N., Wong, Y. W., Ching, P. C., and Chi, H. S. "ISIS: A Multilingual Spoken Dialog System developed with CORBA and KQML Agents". In *Proceedings of 6th International Conference on Spoken Language Processing*, 2000.

[5] Meng, H., Chan. S. F., Wong, Y. F., Chan, C. C, Wong, Y. W., Fung. T. Y., Tsui, W. C., Chen, K. , Wang, L, Wu, T. Y., Li, X. L., Lee, T., Choi, W. N., Ching, P. C., and Chi, H. S. "ISIS: A Learning System with Combined Interaction and Delegation Dialogs". In *Proceedings of Eurospeech*, 2001.

[6] The UCLA Language Materials Projects. "Mandarin Profile".

[7] Lee, T., Meng, H. M., Lau W., Lo, W. K., and Ching, P. C. "Micro-prosodic Control in Cantonese Text-to-Speech Synthesis". In *Proceedings of the 6th European Conference on Speech Communication and Technology, Vol.4, pp.1855 - 8*, September 1999.

[8] Lee, L. S., Tseng, C. Y., and Ouh-Young, M. "The Synthesis Rules in a Chinese Text-to-Speech System". *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(9):pp.1309–1320, 1989.

[9] Linguistic Society of Hong Kong. "Hong Kong Jyut Ping Character Table". Linguistic Society of Hong Kong Press, 1997.

[10] Dudley, H. "A Machine That Talks With the Voice of Man". Science News Letter, pp.19, January 1939.

[11] Fant, G. *"Acoustic Theory of Speech Production"*. The Hague, 1960.

[12] Wilhelms-Tricarico, R. "Physiological Modeling of Speech Production: Methods for Modeling Soft-tissues Articulator". *Journal of the Acoustical Society of America*, 97:3085–3098, 1995.

[13] Flanagan, J. L., Ishizaka, K., and Shipley, K. L. "Synthesis of Speech from a Dynamic Model of the Vocal Cords and Vocal Tract". *Bell System Technology Journal*, 54(3):485–506, 1975.

[14] Rabiner, L. R. and Schafer, R. W. *"Digital Processing of Speech Signals"*, pages 55–105. Prentice-Hall, 1978.

[15] Potter, R. K., Kopp, G. A., and Green, H. C. "Visible Speech". van Nostrand, New York, 1947.

[16] Allen, J., Sharon, M., Hunnicutt, S., and Klatt, D. "From Text to Speech: The MITalk System". *Cambridge University Press, Cambridge, UK*, 1987.

[17] Bickley, C. A., Stevens, K. N., and Williams, D. R. *"A framework for Synthesis of Segments Based on Pseudoarticulatory Parameters"*, pages 211–220. New York, Springer-Verlag, 1997.

[18] Klatt, D. "Review of text-to-speech conversion for English". *Journal of the Acoustical Society of America*, 82(3):737–793, September 1987.

[19] Olive, J. P., Greenwood, A., and Coleman, J. S. *"Acoustics of American English Speech: a Dynamic Approach"*. New York, Springer-Verlag, 1993.

[20] Huang, X., Acero, A, Adcock, J., Hon, H., Goldsmith, J., Liu, J., and Plumpe, M. "Whistler: A Trainable Text-to-Speech System". In *Pro-*

ceedings of International Conference on Spoken Language Processing, pages 2387–2390, Philadephia, 1996.

[21] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and van der Vrecken, O. "The MBROLA Project: Towards a Set of High Quality Speech Synthesizers Free of Use for Commercial Purposes". In *Proceedings of International Conference on Spoken Language Processing*, 1996.

[22] Olive, J. P. "Rule Synthesis of Speech from Dyadic Units". In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 568–570, Hartford, CT, 1977.

[23] Shih, C. L. and Liberman, M. Y. "A Chinese Tone Synthesizer". Technical report, AT&T Bell Laboratories, 1987.

[24] Mao, Y. H, Huang, J, and Zhang, G. Z. "A Chinese Mandarin Speech Output System". *European Conference on Speech Technology*, 2:87–92, 1987.

[25] Chiou, H. B., Wang, H. C., and Chang, Y. C. "Synthesis of Mandarin Speech Based on Hybrid Concatenation". *Computer Processing and Oriental Languages*, Vol.5:pp.217–231, November 1991.

[26] Shih, C. L. and Sproat, R. "Issues in Text-to-Speech Conversion for Mandarin".

[27] Ouh-Young, M. *"A Chinese Text-to-Speech System"*. PhD thesis, National Taiwan University, 1985.

[28] Ouh-Young, M., Shie, C. J., Tseng, C. Y, and Lee, L. S. "A Chinese Text-to-Speech System Based upon a Syllable Concatenation Model". In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 2439–2442, 1986.

[29] Liu, C. S, Tsai, M. F, Hsyu, Y. H., Lu, C. C., and YuS. M. "Yi Xianxing Yuce Bianma Wei Hechengqi de Zhongwen Wenju Fan Yuyin Xitong (A Chinese Text-to-Speech Based on LPC Synthesizer)". *TL Technical Journal*, 19(3,):269–285, 1989.

[30] Liu, C. S., Ju, G. H., Wang, W. J., Wang, H. C., and Lai, W. H. "A New Speech Synthesizer for Text-to-Speech System Using Multipulse Excitation with Pitch Predictor". In *Proceedings of International Conference on Computer Processing of Chinese and Oriental Languages*, pages 205–209, 1991.

[31] Ju, G. H., Wang, W. J., Liu, C. S., and Lai, W. H. "Yi Tao Yi Duomaichong Jifa Yuyin Bianmaqi Wei Jiagou zhi Jishi Zhongwen Wenju Fan Yuyin Xitong (A Real-time Implementation of Chinese Text-to-Speech System Based on the Multi-pulse Excited Speech Coder)". *TL Technical Journal*, 21(4):431–440, 1991.

[32] Chan, N. C. and Chan, C. K. "prosodic rules for connected mandarin synthesis". *Journal of Information Science and Engineering*, 8:261–281, 1992.

[33] Cai, L. H., Liu, H., and Zhou, Q. F. "Design and Archievement of a Chinese Text-to-Speech System under Windows". *Microcomputer*, 3, 1995.

[34] Chu, M. and Lu, Shinan. "High Intelligibility and Naturalness Chinese TTS System and Prosodic Rules". In *Proceedings of the XIII International Congress of Phonetic Sciences*, pages 334–337, Stockholm, 1995.

[35] Hwang, S. H., Wang, Y. R., and Chen, S. H. "A Mandarin Text-to-Speech System". In *"Proceedings of International Conference on Spoken Lanugage Processing"*, 1996.

[36] Chu, M., Tang, D., Si, H., Tian, X, and Lu, S. "Research on Perception of Juncture Between Syllables in Chinese". *Chinese Journal of Acoustics*, 17(2):143–152, 1998.

[37] Chu, M., Peng, H., Yang, H. Y., and Chang, Eric. "Selecting Non-Uniform Units from a Very Large Corpus For Concatenative Speech Synthesizer". In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2001.

[38] Sagisaka, Y. "Speech Synthesis by Rule Using an Optimal Selection of Non-Uniform Synthesis Units". In *Proceedings of International Conference on Acoustics, Speech and Signal Proceeding"*, pages 679–682, New York, 1988.

[39] Fung, T. Y. and Meng, H. M. "Concatenating Syllables for Response Generation in Domain-Specific Applications". In *Proceedings of International Conference on Acoustics, Speech and Signal Proceeding*, 2000.

[40] Kenney Ng. "Survey of Data-Driven Approaches to Speech Synthesis". Massachusetts Institute of Technology, October 1998.

[41] Hon, H., Acero, A., Liu, J. Huang, X., and Plumpe, M. "Automatic Generation of Synthesis Units from Trainable Text-to-Speech Systems". In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 293–306, 1998.

[42] Campbell, N. and Black, A. W. *"Progress in Speech Synthesis"*, chapter Prosody and the Selection of Source Units for Concatenative Synthesis, pages 279–292. Springer Verlag, 1997.

[43] Black, A. W. and Lenzo, K. A. *"Building Voices in the Festival Speech Synthesis System*, chapter 10: Limited Domain Synthesis. Speech Group at Carnegie Mellon University, http://www.cstr.ed.ac.uk/projects/festival/docs/festvox/, July 2000.

[44] Black, A. W. and Campbell, N. "Optimising Selection of Units from Speech Databases for Concatenative Synthesis". In *Proceedings of Eurospeech*, pages 581–584, Madrid, Spain, October 1995.

[45] Campbell, N. "CHATR: A High-Definition Speech Re-sequencing Sys-

tem". Acoustic Society of America and Acoustical Society of Japanm, Third Joint Meeting, December 1996.

[46] Hunt, A. and Black, A. W. "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database". In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 373–376, Atlanta, 1996.

[47] Takeda, K., Abe, K., and Sagisaka, Y. *"Talking Machine: Theories, Models, and Designs"*, chapter 1, On the Basic Scheme and Algorithms in Non-Uniform Unit Speech Synthesis, pages 93–105. Elsevier Science, Holland, 1992.

[48] Chou, F. C., Tseng, C. Y., and Lee, L. S. "Selection of Waveform Units For Corpus-Based Mandarin Speech Synthesis Based on Decision Trees and Prosodic Modification Costs". In *Proceedins of Eurospeech*, 1999.

[49] Huang, X. D., Acero, A., and Hon, H. W. *"Spoken Language Processing: A Guide to Theory, Algorithm and System Development"*. Prentice Hall, 2001.

[50] Moulines, E. and Charpentier, F. "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphone". *Speech Communication*, 9(5):453–467, 1990.

[51] Moulines, E. and Verhelst, W. *"Speech Coding and Synthesis"*. Elsevier, 1995.

[52] Hamon, C., Moulines, E., and Charpentier, F. "A Diphone Synthesis System Based on Time-Domain Prosodic Modifications of Speech". In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 238–241, Glasgow, Scotland, May 1989.

[53] Chu, M and Ching, P. C. "A Cantonese Synthesizer Based on TD-PSOLA Method". In *Proceedings of International Symposium on Mixing in Industrial Processes (ISMIP)*, pages 262–267, Taipei, 1997.

[54] Beckman, M. E. and Ayers, G. M. "Guidlines for ToBI Labelling". Technical report, Ohio State University, 1993.

[55] Lee, L. S., Tseng, C. Y. Fellow IEEE, and Hsieh, C. J. "Improved Tone Concatenation Rules in a Formant-Based Chinese Text-to-Speech System". *IEEE Transactions on Speech and Audio Processing*, 1(3), 1993.

[56] Shih, C. L. "The Phonetics of the Chinese Tonal System". Technical report, AT&T Bell Laboratories, 1987.

[57] Shih, C. L. and Kochanshi, G. P. "Chinese Tone Modeling with Stem-ML". In *Proceedings of International Conference on Spoken Language Processing*, 2000.

[58] Lucent Technologies Bell Labs Innovations. Bell labs text-to-speech synthesis overview.

[59] Taylor, P. et al. "The Architecture of the Festival Speech Synthesis System". In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, 147-151.

[60] Tayor, P. A. "Concept-to-Speech Synthesis by Phonological Structure Matching". *Article submitted to Royal Society*, 1999.

[61] Yi, J. "Time-Domain PSOLA Concatenative Speech Synthesis Using Diphones".

[62] Lo, W. K. and Ching, P. C. "Phone-based Speech Synthesis with Neutral Network and Articulatory Control". In *Proceedings of Internation Conference on Spoken Language Processing*, 1996.

[63] Chu, M. and Ching, P. C. "A Hybrid Approach to Synthesize High Quality Cantonese Speech". In *Proceedings of Acoustics, Speech and Signal Processing*, 1998.