GENERATION OF PROSODY AND SPEECH FOR MANDARIN CHINESE

DONG MINGHUI

(BS, University of Science and Technology of China, 1992)

(MS, Peking University, 1995)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

SCHOOL OF COMPUTING

NATIONAL UNIVERSITY OF SINGAPORE

2002

Acknowledgments

The completion of this thesis would not have been possible without the help of many people to whom I would like to express my heartfelt appreciation.

I would like to express my deepest gratitude to my supervisor, Dr. Lua Kim Teng, who has always been helping me in both my research and my life. He has always been encouraging me to do my best when I encounter difficulties. This work would not have been possible without his guidance.

I thank National University of Singapore and School of Computing for providing me a pleasant working environment. I also would like to thank every member in the Computational Linguistics Laboratory for all the help during the years of my study.

I thank InfoTalk Technology for putting me on the frontier of TTS technology, for giving me chances to investigate the problems and to apply what I have learned in various aspects of TTS system.

Thanks are also given to the reviewers of my thesis for their valuable comments, which help to improve this thesis. Special thanks go to Dr. Li Haizhou for reviewing and commenting my thesis. Thank Miss Ma. Ledda T. Santiago for proofreading my English writing.

Finally, the greatest gratitude goes to my parents, wife, brothers, sister, and my little son for supporting me and encouraging me in all the years.

Table of Contents

ACKNOWLEDGMENTS I
TABLE OF CONTENTS II
SUMMARY VII
LIST OF TABLES VIII
LIST OF FIGURES X
CHAPTER 1 INTRODUCTION 1
1.1 Knowledge of TTS 1
1.1.1 Text-to-Speech 1
1.1.2 Prosody 3
1.1.3 Speech Synthesis by Unit Selection 4
1.2 Research Overview 5
1.2.1 Problem Statement 5
1.2.2 Brief Description of the Work 8
1.2.3 Problems not Concerned in the Work 9
1.3 Outline of the Thesis 11
CHAPTER 2 FOUNDATIONS 12
2.1 Basics of Chinese 12
2.1.1 Words 12
2.1.2 Phonetics of Chinese 13
2.1.3 Mandarin 14
2.2 Chinese Prosody 14
2.2.1 Tone 14
2.2.2 Intonation Theory of Chinese 16
2.2.3 Rhythm 16
2.3 Classification and Regression Tree (CART) 17
2.3.1 Classification Tree or Regression Tree 17
2.3.2 Splitting Criteria 20
2.3.3 Building Better Tree 21
2.4 Formulas 22
2.4.1 Mutual Information 22
2.4.2 Pearson Product Moment Correlation Coefficient 22

CHAPTER 3 SPEECH CORPUS CONSTRUCTION 23

3.1 Speech Corpus Construction and Processing 23

3.1.1 Consideration of Number of Speakers 23

3.1.2 Speech Data 24

3.1.3 Text Data 25

3.1.4 Data Attributes 26

3.2 Phonetic Statistics of Chinese 28

3.2.1 Context Independent Unit 29

3.2.2 Context Dependent Unit 30

3.2.3 Grouping Context Units by Initial and Final 32

3.2.4 Considering Loose Coarticulation 33

3.2.5 Unit Distribution for Different Context Considerations 34

3.3 Corpus Evaluation 35

3.3.1 Word Frequency 36

3.3.2 Syllable Coverage 36

3.3.3 Statistics 37

3.3.4 Conclusion 39

3.4 Summary 39

CHAPTER 4 PROSODIC BREAK PREDICTION 40

4.1 Introduction 40

4.1.1 Prosodic Break 40

4.1.2 Review of Existing Approaches 41

4.1.3 Review of Work for Chinese 43

4.2 Determination of Prosodic Breaks 44

4.2.1 Chinese Prosodic Structure 44

4.2.2 Issues of Prosodic Break in this Work 46

4.3 Prosodic Word Detection 48

4.3.1 Prosodic Word 49

4.3.2 Patterns of Prosodic Words 50

4.3.3 Baseline Model 53

4.3.4 Grouping POS Categories 53

4.3.5 Single Word Categories 54

4.3.6 Dependency on Previous Break 54

4.3.7 Global Optimization 55
4.3.8 Experiments 58
4.4 Minor Phrase Break Detection 63
4.4.1 CART Approach 66
4.4.2 Dependency Model 66
4.4.3 Experiments 68
4.5 Discussion 72
4.6 Summary 73
CHAPTER 5 PROSODY PARAMETERS 74
5.1 Introduction 74
5.1.1 Pitch Contour 75
5.1.2 Duration 76
5.1.3 Energy 77
5.1.4 Previous Approaches for Chinese Prosody 78
5.2 Problems and Solutions 78
5.2.1 Problems of Prosody for Unit Selection 79
5.2.2 Implementation of Perceptual Effects 80
5.2.3 Solutions for the Problems 83
5.3 Prosody Parameters for Unit Selection 84
5.3.1 Duration and Energy 84
5.3.2 Pitch Contour 88
5.3.3 Candidate Prosody Parameters 91
5.4 Parameter Determination 92
5.4.1 Parameter Evaluation 92
5.4.2 Parameter Selection 93
5.5 Prediction of Prosody 94
5.5.1 Features for Prediction 94
5.5.2 Prediction Ability of Features 96
5.5.3 Prediction Model 98
5.6 Experiments 98
5.6.1 Parameter Determination 98
5.6.2 Single Feature in Prediction 112

5.6.3 Combined Features for Prediction 121
5.6.4 Prediction of All Parameters 126
5.7 Summary 128
CHAPTER 6 UNIT SELECTION WITH PROSODY 130
6.1 Introduction 130
6.1.1 Unit Selection-Based Synthesis 130
6.1.2 Problems of Prosody in Unit Selection 134
6.2 Unit Selection Model in this Work 135
6.2.1 Unit Specifications 135
6.2.2 Corpus Coverage 136
6.2.3 Implementation of Prosody by Unit Selection 137
6.2.4 Costs for Unit Selection 137
6.2.5 Dynamic Programming 139
6.3 Definition of the Cost Function 141
6.3.1 Phonetic Cost of Unit ($C_{Phonetic}$) 141
6.3.2 Prosodic Cost of Unit ($C_{Prosodic}$) 143
6.3.3 Smoothness Cost between Two Units (C_{Smooth}) 145
6.3.4 Connection Importance Factor Between Two Units (I_{Conn}) 147
6.3.5 Total Cost 147
6.3.6 Weight Determination 148
6.4 Summary 150
CHAPTER 7 EVALUATION 151
7.1 Introduction of Speech Quality Evaluation 151
7.1.1 Segmental Unit Test 151
7.1.2 Sentence Level Test 152
7.1.3 Overall Test 153
7.1.4 Objective Evaluation 153
7.2 Evaluation of Speech Quality 154
7.2.1 Testing Problem of this Work 154
7.2.2 Evaluation Methods in this Work 155
7.2.3 Testing Material Selection 158
7.3 Experiments 159
7.3.1 Testing Text Selection 159

7.3.2 Parametric Prosody vs Symbolic Prosody 160
7.3.3 Break and Tone Accuracy 163
7.3.4 Quality of Synthetic Speech 165
7.3.5 Speed of TTS system 168
7.4 Discussion 171
7.5 Summary 173
CHAPTER 8 CONCLUSION 174
8.1 Summary of the Research 174
8.2 Contributions 175
8.3 Future Work 177
BIBLIOGRAPHY 179
APPENDIX 191
A. Part-of–speech Tag Set of Peking (Beijing) University 191
B. Features for Unit in Speech Inventory 192
C. Sentences for Listening Testing 193
D. Text Example for Intelligibility Testing 195
E. List of Published Papers 196

Summary

This research is an investigation of the problem of prosody generation for Mandarin Chinese text-to-speech system. I mainly work on two issues of prosody: (1) The prediction of prosodic phrase breaks, especially the prediction of prosodic word break. (2) The design, evaluation, and selection of prosody parameters for unit selection based synthesis.

This work uses a speech corpus read by a female professional speaker. During the evaluation of speech corpus, the problem of speech unit distribution of Chinese language is first investigated. The speech corpus is then evaluated to find if it is suitable for this work.

The problem of prosodic break has been investigated. The factors that affect the performance of prosodic break are examined. Dependency models for break prediction are developed. The experiments show that the models produce better result than the simple CART approach.

The approaches of designing, evaluating, and selecting prosody parameters are given. Some prosody parameters are defined to suit the nature of Chinese speech and the approach of unit selection. The parameters defined in this work are intended to overcome the major speech problems in speech synthesis. We highlight the problems of correctly representing perceptual prosody information in this work. The defined parameters are examined from statistical views and recognition views. A clustering approach is used to remove redundancy in prosody parameter definition. The relationship between the parameters and features for prediction has been investigated.

In the unit selection-based synthesis, the defined parametric prosody expression is applied in cost function. Some experiments are designed to better evaluate the system. The experiments show that the use of parametric prosody representation significantly improved the quality of speech.

List of Tables

Table 1.1 Tasks of this work....9
Table 2.1 Initials and Finals in Chinese....13
Table 3.1 Data tiers of the corpus....27
Table 3.2 Example of text tiers in corpus....28
Table 3.3 Class of right edge (final) of syllable....32
Table 3.4 Class of left edge (initial or final for null-initial syllable) of syllable....33
Table 3.5 Classification of initials for tightness of connection. 34
Table 3.6 Number of units for coverage of context dependent units....35
Table 3.7 Coverage of context dependent units of the corpus....36
Table 3.8 Number of text units and prosodic units in the corpus....37
Table 3.9 Length distribution of words in the corpus....37
Table 3.10 Frequency of POS in corpus....38
Table 3.11 Occurrence distribution of toneless syllable in the corpus....38
Table 3.12 Distribution of tones in the corpus....38
Table 4.1 Prosodic word patterns in terms of POS....51
Table 4.2 Prosodic word patterns in terms of word length....51
Table 4.3 Mutual information between break type and features....52
Table 4.4 Accuracy of using different feature sets....60
Table 4.5 Accuracy of different word group size....61
Table 4.6 Performance comparison for CART approach and Dependency model....62
Table 4.7 Speed comparison for CART approach and Dependency model for prosodic word break prediction....63
Table 4.8 Mutual information between break type and previous break type for minor phrase....65
Table 4.9 Mutual information between break type and previous and next POS types for minor phrase....65
Table 4.10 Result of break prediction using CART and POS sequence....69
Table 4.11 Result of break prediction using dependency model....69
Table 4.12 Speed comparison for CART approach and Dependency model for phrase break prediction....72
Table 5.1 Accuracy for tone recognition....101
Table 5.2 Correlation values between parameters for tone....102
Table 5.3 Recognition result of StartOfPW....105

Table 5.4 Correlation values between break related variables 107
Table 5.5 Final clusters in parameter clustering 110
Table 5.6 Correlation values between selected parameters 110
Table 5.7 Comparison of factors determining pitch mean 113
Table 5.8 Comparison of factors determining duration 116
Table 5.9 Comparison of factors determining Energy 119
Table 5.10 Stepwise training for PitchMean 121
Table 5.11 Stepwise training for Duration 123
Table 5.12 Stepwise training for Energy 124
Table 5.13 Result of the prosody parameter prediction 127
Table 6.1 Final weights in the cost function 150
Table 7.1 MOS scores for listening test. 157
Table 7.2 Methods used in cost test 161
Table 7.3 Result of rate of inappropriate units(RIU) 161
Table 7.4 Accuracy of break in speech 164
Table 7.5 Result of correctly implemented tones 165
Table 7.6 Result for intelligibility test (Rate of recognized units) 167
Table 7.7 Result for naturalness test 168
Table 7.8 Speed of unit selection dependent on beam width 169
Table 7.9 Synthesis speed comparison 170
Table 7.10 Time breakdown for TTS 171

List of Figures

Figure 1.1 Typical Framework of a TTS System 2
Figure 2.1 Decomposition of a Chinese base syllable 13
Figure 2.2 Tones and pitch tracks of base syllable “ma” (Xu, 1997) 15
Figure 2.3 Example of classification tree (Answer “yes” to left, “no” to right child) 18
Figure 2.4 Example of regressin tree (Answer “yes” to left, “no” to right child) 19
Figure 3.1 Example of Chinese prosodic structure 26
Figure 3.2 Example of speech tiers in the corpus (waveform, F0 contour and syllable labels) 27
Figure 3.3 Accumulative coverage of syllables in text corpus. 30
Figure 3.4 Accumulative coverage of pinyin trigram 31
Figure 3.5 Accumulative coverage of syllable with context considered 31
Figure 4.1 Prediction of probability using Classification tree 56
Figure 4.2 Distribution of number of syllables in phrase 64
Figure 4.3 Calculation of probability using CART 66
Figure 4.4 Calculation of probability using CART in dependency model 67
Figure 4.5. Comparison of precision values for phrase break prediction using the CART and dependency model 70
Figure 4.6. Comparison of recall values for phrase break prediction using the CART and dependency model 70
Figure 5.1 Prediction of prosody 81
Figure 5.2 Syllable duration normalization 85
Figure 5.3 Illustration of pitch curves of tone 89
Figure 5.4 Illustration of prosody parameters 90
Figure 5.5 Boxplots for PitchMean by tone type 99
Figure 5.6 Boxplot for PitchRange by tone type 100
Figure 5.7 Boxplots for PitchStart by tone type 100
Figure 5.8 Boxplots for PitchEnd by tone type 100
Figure 5.9 Boxplots of Duration by boundary type 103
Figure 5.10 Boxplots of EnergyStart by boundary type 104
Figure 5.11 Boxplots of EnergyHalfPoint by boundary type 104
Figure 5.12 Boxplots of EnergyEnd by boundary type 105
Figure 5.13 Dendrogram for clustering parameters 108

Figure 5.14 Similarity level in paramter clustering step....109
Figure 5.15 Stepwise training of PitchMean....122
Figure 5.16 Stepwise training of Duration....123
Figure 5.17 Stepwise training of Energy....125
Figure 5.18 EnergyRMS changing with location of syllable in utterance....126
Figure 6.1 Illustration of unit selection....133
Figure 6.2 Illustration of unit cost calculation....138
Figure 6.3 Direct calculation of connection cost....138
Figure 6.4 Indirect calculation of connection cost....139
Figure 6.5 Connection cost calculation....146
Figure 7.1 Text selection for listening test....160
Figure 7.2 Speed of unit selection....169
Figure 7.3 Time breakdown of the TTS....170

Chapter 1 Introduction

The aim of this research is to develop an approach to generate good prosody from Mandarin Chinese text and then apply the prosody to a speech generation component (synthesizer) to generate high quality speech. Specifically, we investigate what prosody description is suitable for unit selection based synthesis approach.

The research is carried out through building a full size Chinese text-to-speech system, which is used as a test bed for studying and evaluating algorithms and approaches.

1.1 Knowledge of TTS

In order to explain the work of this research, in this section, we introduce some of the topics related to the research.

1.1.1 Text-to-Speech

Text-to-speech synthesis (TTS) is the automatic conversion of any plain text to speech (Shih and Sproat, 1996). The generated speech is expected to resemble that of a native speaker of the language as closely as possible. The input text usually exists in machine-readable form, such as a text file. The subject in this research is Mandarin Chinese TTS. Therefore, the input of the system is Chinese text in the form of Chinese codes (such as GBK for Simplified Chinese or Big5 for Traditional Chinese), which can be in a text file format, and the output of the system is speech signal, which may be stored in a computer as a waveform file.

In the past decades, much progress has been made in Chinese TTS systems and many systems have been built (Lee et al., 1989,1993; Chan et al. 1992; Chen et al., 1998; Shih and Sproat, 1996; Chou and Tseng, 1998). Like TTS systems in other languages, a typical TTS system consists of three main parts, which are text analysis,

prosody generation, and speech signal synthesis. Figure 1.1 shows a typical framework of a TTS system.

The input of a TTS system is usually raw text. Text analysis is to change the raw text into the format that prosody generation and synthesis parts can accept. The raw text may consist of non-Chinese characters (symbols, digits, etc). Before doing other things, a text normalization process converts them into Chinese text. After normalization, the text becomes a sequence of Chinese characters. As there is no space delimiter between words in Chinese, to perform further analysis, words should be extracted from the sentence. Word segmentation identifies words in the continuous Chinese text. Moreover, POS (Part-of-speech) is one of the basic information for understanding a sentence. POS tagging process classifies each word into a category. POS information may be useful in analysis of prosody structure, as will be shown in later chapters. Another task of text analysis is to convert the Chinese text into phonetic representations for producing correct sounds in the generated speech.

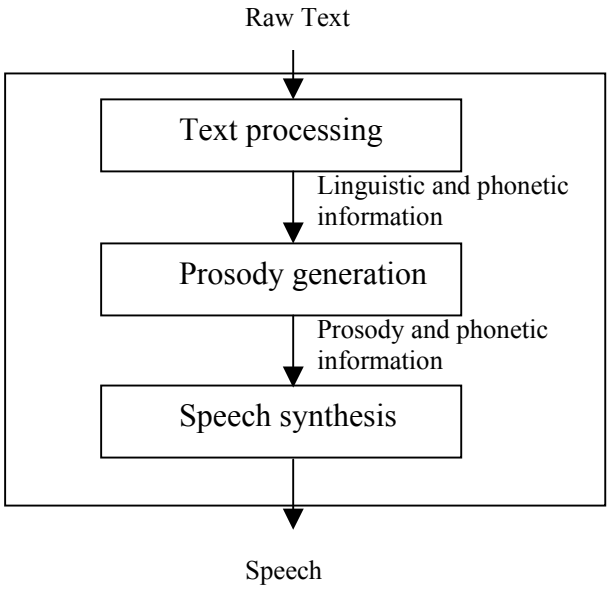


Figure 1.1 Typical Framework of a TTS System

The second part of a TTS system is prosody generation. Proper prosody should be generated according to the linguistic and phonetic information contained in the sentence. The prosody includes rhythm, pause, accent, pitch, duration, and other perceptually identifiable acoustic features in speech. The process of prosody generation usually does the following work:

- **Determining Symbolic Representation of Prosody:** Usually, several levels of break are defined to give a prosody structure of a sentence. The breaks will determine the duration of pause between words and will affect prosody parameters, such as duration of speech units, pitch contour, etc. In some languages (e.g. English), labels for stress, accent and boundary tone also need to be determined at this stage. The breaks and labels are symbolic representations that describe some abstract prosody events.

- **Determining Parametric Representation of Prosody:** Prosody parameters are a set of quantitative parameters that represent prosody (pitch contour, duration, and energy) of the utterance to be generated. These parametric representations are continuous values that measure the acoustic properties of speech. A model is usually built to convert all the available symbolic information (linguistic and phonetic inputs, prosodic breaks, and intermediate labels) into some desired parameters.

The third part of a TTS system is the synthesis component, which transforms the pronunciation and prosody information into speech signal. The segmental (linguistic) and supra-segmental (prosody) information should be well presented in the generated speech. The pronunciation is usually done by selecting the correct synthesis unit, while the realization of prosody is either by transformation of the synthesis units or by selecting the proper units that match the target prosody.

1.1.2 Prosody

The ultimate goal of a TTS system is to make the system read text like a human. The naturalness of speech depends on how much acoustic information of natural speech is contained in the reconstructed speech. Natural human speech usually contains two

different sorts of information: segmental information and suprasegmental information. The segmental information refers to what the speaker says. The suprasegmental information refers to how the speaker says. Same segmental information with different supra-segmental information may result in different meanings. For example, “Good.” and “Good?” have the same segmental information but different intonations, resulting in different meanings.

Suprasegmental information is usually referred to as prosody in literature. Prosody generally consists of certain properties of the speech signal such as audible changes in pitch, loudness, syllable length, pause, and so on. Perceptually, prosody is usually perceived as break, tone, accent, intonation, etc. Acoustically, prosody is measured by fundamental frequency (F0) contour of speech waveform, length of duration, and energy level of speech units, etc.

Fundamental frequency is usually regarded as the most important element of prosody. As fundamental frequency is perceptually identified as pitch, in many literatures, it is referred to as pitch. In this work, we use the term “pitch” to mean fundamental frequency in most occasions. We use pitch contour to mean funamental frequency contour, which is also referred to as intonation contour in some literatures.

1.1.3 Speech Synthesis by Unit Selection

There has been a lot of research on speech synthesis in the past decades. All the methods can be classified into three major categories (Flanagan, 1972), which are articulatory synthesis, formant synthesis, and concatenation synthesis. Articulatory synthesis attempts to model the human speech production systems, while formant synthesis and concatenation synthesis attempt to only model resultant speech. Formant synthesis generates speech with the support of a database of rules. Concatenation synthesis concatenates pre-recorded speech units to form the final speech. During the synthesis process, the units are usually changed to fit the prosody requirements.

Most of the traditional speech synthesis approaches use signal-processing techniques to construct or transform speech signals during synthesis process. This

usually generates speech with a machine-like voice. As the development of hardware, computer has more memory and more powerful computation power. It becomes more realistic to store as many speech units as possible. Therefore, an extreme approach emerged. The approach uses a huge prerecorded corpus (Black and Campbell, 1995; Hunt and Blank 1996). During synthesis, we only need to select the best synthesis units and then concatenate them without any modification. As there is no signal processing to the original speech signal, the synthetic speech can be very natural.

1.2 Research Overview

1.2.1 Problem Statement

As we have stated, speech contains two kinds of information, which are segmental information and suprasegmental information (prosody). Segmental information determines the intelligibility of speech, while suprasegmental information determines the naturalness of speech. The aim of this work is to generate high quality speech. To generate high quality speech, we need to generate speech with proper segmental information and proper suprasegmental information (prosody).

Unit selection based approach is considered a way to improve the segmental information for synthetic speech. Since speech pieces are directly copied to final speech during synthesis process, the generated speech can keep the segmental information as much as possible.

When we decide to use unit selection based approach for synthesis, the main problem of generating high quality speech becomes the generation of natural prosody. To generate natural prosody, we have to (1) generate a correct prosodic structure and (2) generate a proper representation of prosody.

In Chinese, syllables are usually grouped into prosodic words. Prosodic words are further grouped together to form prosodic phrase. The existence of prosodic structure makes speech natural. To synthesize speech with a correct prosodic structure, we have to investigate the problems of the placement of prosodic breaks, especially the prosodic word breaks.

For unit selection based approach, it is a problem to ensure that the suprasegmental information of synthetic speech is correct and the best. Unlike other approaches, the unit selection based approach is a pattern matching process, in which prosody of speech unit cannot be changed. We may have the following problems in dealing with this. (1) How to measure the mismatch between target unit and selected unit? (2) What representation is needed for describing prosody of units? (3) How to keep the parameter set concise but sufficient? (4) What factors are important in predicting prosody parameters?

To investigate the problems of prosodic break and prosody parameters, we also need a reliable speech corpus and reliable evaluation approaches. Therefore, the main problems to be solved in this work can be described from the following aspects:

(1) Corpus Evaluation

Both corpus-based prosody generation and unit selection-based speech synthesis approaches require speech corpora. To better investigate the prosody and synthesis problems, the speech corpus should be well designed to have a good coverage of the prosody and speech phenomena. Due to the large number of unit combinations in Chinese, it is a big challenge to design an inventory that covers prosody phenomena as largely as possible, yet to keep the size of the inventory as small as possible. The distribution of units in this language should be investigated. The speech corpus for this work should be well evaluated before it is used.

(2) Prosodic Break Prediction

One of the most important aspects of Chinese prosody is the organization of speech units when speaking. Linguists have found that there is a hierarchical structure for Chinese prosody. Syllables are grouped together to form prosodic groups. Due to the existence of different levels of prosodic group, listeners can perceive different types of prosodic break. The breaks make listener to understand speech better. However, this hierarchical structure cannot be well used in Chinese TTS system due to poor prediction approaches. Especially, we need to investigate the approaches and factors in the prediction of prosodic words.

(3) Prosody Parameter Design and Prediction

There were some prosody models designed for Chinese (refer to 5.1.4). However, they have the following shortcomings:

(1) They are designed for signal processing based synthesis (e.g. PSOLA, etc), in which signals are transformed according to prosody requirements. They are normally unsuitable for unit selection. There is no pitch contour mismatch between units in signal processing based synthesis. However, it is a problem to measure a prosody mismatch during unit selection-based synthesis process.

(2) The general prosody parameters (duration, energy, and pitch contour) cannot capture all the important aspects of prosody. For example, duration analysis showed that boundary units (e.g. start and end units of a prosodic word or a phrase) have longer durations than other units. However, if we select a long unit only based on duration, the selected unit is not necessarily a unit that we expect. Duration alone cannot distinguish boundary units from non-boundary ones, which however are quite different in perception. Therefore, some more prosody parameters should be investigated to account for these prosody differences in units. Another important aspect for Chinese prosody is tone. How to effectively express tone information is also a problem.

(3) When we define many parameters to account for different aspects of prosody, the defined parameters may have redundancy. How to select a small set of parameters yet to describe the main prosody properties is a problem.

(4) To understand the problem of prosody prediction, we need to further investigate the relationship between the parameters and the features.

(4) Unit Selection with Prosody

Unit selection based approach has been used by English and other languages. However, integration of prosody in unit selection remains a problem. Some systems (e.g. Chu et al, 2000) integrate symbolic representation of prosody in their work. Symbolic representations are discrete values to describe prosody events, such as break

types, accent marks etc. The symbolic representations can capture some of the prosody differences. However, the discrete values cannot provide an accurate distinction between units. Hence, the best units may not be selected due to the absence of proper distinction measures. Some work tried to use parametric parameters (e.g. Campbell et al, 1996), however, the parameters are not carefully designed for unit selection based approach and the way to apply the prosody is not well considered. For example, variation of prosody parameters was not well handled in their work.

Evaluation of synthetic speech is always problematic for two reasons: (1) Language is an infinity set. Complete testing is impossible. (2) Speech quality is often evaluated by human perception. Thus, evaluation is difficult to be conducted.

To have a fair evaluation of speech, the testing material and testing approach is very important. Designing text that has a good coverage of the language in question should be investigated. To better evaluate the performance of the defined prosody parameters using subjective test, proper testing approach should be used.

1.2.2 Brief Description of the Work

This work is to investigate the problem of the prediction of prosodic breaks and prosody parameters. Especially, we want to investigate how prosody is designed, predicted, and applied in the unit selection based synthesis. To achieve this goal, we have to work on four main tasks. The four main tasks are as shown in Table 1.1.

The first part is corpus preparation. We will build a good corpus for our main research in this part. In addition, we will evaluate the corpus to make sure it is suitable for this work.

The second part is prosodic break prediction. In prosodic break prediction, we will propose models for predicting the breaks. We will investigate the factors for the prediction of prosodic words.

The third part is the determination and prediction of prosody parameters. In prosody parameter determination, we will propose an approach to decide what kind of prosody description should be used for the unit selection based approach. Especially,

we will propose an approach to convey the tone and break information in the parameters. We will remove the redundancy of the parameters.

The fourth part is the unit selection with prosody. In this part, we will integrate prosody parameters into cost function to help unit selection. We will also design testing texts and testing approaches for listening test.

Tasks	Subtasks
Corpus preparation	Constructing corpus
	Analyzing distribution of Chinese units
	Evaluating the corpus
Prosodic break prediction	Analysizing prosodic words
	Proposing model for prosodic word prediction
	Proposing model for minor prosodic phrase prediction
Prosodic parameter determination	Defining prosodic parameter
	Evaluating prosodic parameters
	Selecting prosodic parameters
	Analyzing prediction factors
Unit selection with prosody	Defining cost functions
	Designing testing text
	Evaluating synthetic speech

Table 1.1 Tasks of this work

1.2.3 Problems not Concerned in the Work

To better understand and avoid misunderstanding of the scope of this work, we list some issues that may be raised.

(1) Speaker Dependent or Speaker Independent

The work is about text-to-speech system. The synthetic speech should come from only one speaker. To make the generated speech resemble the voice and the speaking style of the original speaker, the prosody model should also be built from the same speech data. Therefore, the TTS system is a speaker dependent system.

Different speakers may have different prosody styles, such as the habits of breaking within a sentence. However, since we are going to generate prosody for TTS system, this research deals with common prosody characteristics among general native speakers. Prosody differences among speakers are not the main issue of this work.

The speech corpus in the work is read by a speaker with common speaking style. The results produced by the models using the corpus may be speaker dependent. However, the approaches adopted are speaker independent because they are not based on speaker dependent features.

(2) Locality

The speech to be generated is standard mandarin Chinese speech. (Refer to Section 2.1.3) Other dialects are not concerned in this work. To concentrate on TTS, we do not take dialects or locality as part of the work.

(3) Prosody and Emotion

Emotion is one of the expressing forms of prosody. Emotional speech usually has special duration, pitch contour, and energy variation. However, emotion is not the topic of this research. The main aim of this work is to generate speech with general speaking style and voice quality. The generated speech is to be used for general purpose rather than in specific domain or for special use.

(4) Meanings of Prosody

In life, we generally use prosody to mean poem style text. Speech with prosody usually means speech with regular rhythm. However, in the context of text-to-speech synthesis, prosody means some particular perceptual properties of speech. The prosody in this work means the later. Therefore, any speech segment has its prosody, no matter it has a regular rhythm or not. The meaning of poem style structure of speech is not the part of this work.

1.3 Outline of the Thesis

Chapter 2 introduces the background related to this research. Some basic knowledge of Chinese and Chinese prosody is briefly covered. The training approach CART is briefly introduced.

Chapter 3 describes corpus preparation. The process of generating the corpus is described. The distribution of units in Chinese language is studied. The speech corpus is evaluated also.

Chapter 4 studies the prediction of prosody structure. The problem of prosodic word is first studied. Models for the prediction are given. Some aspects related to the performance are discussed. The problem of minor phrase prediction is also investigated.

Chapter 5 covers prosody parameters for unit selection based synthesis approach. This chapter proposes approaches for designing, evaluating and selecting prosody parameters for unit selection. Prosody parameters are defined. The prosody parameters for describing perceptual prosody effects are evaluated. An approach for selecting parameters is proposed. The relationship between features and parameters is analyzed.

Chapter 6 covers the unit selection-based speech synthesis. The prosody parameters are integrated into unit selection. The cost function for unit selection is defined. The algorithm for unit selection is given. The weights of subcosts are determined.

Chapter 7 describes the evaluation of speech quality. The texts for testing are designed. The performance of the prosody and the TTS system is tested.

Chapter 8 gives a summarization.

Chapter 2 Foundations

In this chapter, some basic knowledge of Chinese and the research findings of Chinese prosody are first covered. Then the main learning approach, CART (classification and regression tree), is described.

2.1 Basics of Chinese

2.1.1 Words

Chinese language differs from Western languages in a number of ways. Chinese is an ideograph language, whose character set is not a closed one. The number of basic Chinese characters is large, ranging from thousands of frequently used characters (GB code) to some twenty thousand ones in a more complete Chinese character code standard (such as GBK or Unicode). A typical system that uses the GB set includes 6763 simplified Chinese characters.

In Chinese, a word is a unit consisting of one or more characters. Most of Chinese words consist of 1 to 4 characters. As there is no generally accepted definition of word, the number of words is not fixed either. Word is defined differently in different applications. A big dictionary may contain 60,000 or even 100,000 Chinese words. As there are always newly generated words, such as compound words and proper names, it is not possible to completely include all possible words in a dictionary.

Another difference between Chinese and Western languages is that there is no space between words in a text of Chinese. Therefore, before the understanding of a sentence, words need to be identified first from a continuous text string of a sentence.

2.1.2 Phonetics of Chinese

Phonetically, each Chinese character is a tonal monosyllable (with exception that around 10 characters have disyllabic pronunciations). Although the number of the characters is large, the number of syllable pronunciations is much less. There are around 408 different syllables in Mandarin Chinese regardless of tone (Chao 1968). Tone is one of the distinguishing characteristics of Chinese. There are five tones for the pronunciation of syllables. Same pronunciation with different tones usually conveys different meanings. There are around 1300 different meaningful pronunciations in Chinese Mandarin if tones are considered. Therefore, usually many Chinese characters share the same pronunciation. It is also possible that one character has more than one pronunciation while having different meanings.

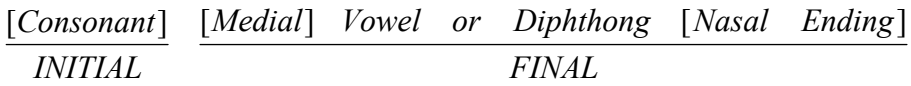


Figure 2.1 Decomposition of a Chinese base syllable

22 INITIALS	B C CH D F G H J K L M N P Q R S SH T X Z ZH NULL-INITIAL
38 FINALS	A AI AN ANG AO E EI EN ENG ER I IA IAN IANG IAO IE IN ING IONG IU IZ IZH ONG OU U UA UAI UAN UANG UENG UI UN UO V VAN VE VN

Table 2.1 Initials and Finals in Chinese

As shown in Figure 2.1 (Chao 1968), conventionally, each Chinese base syllable can be decomposed into an initial-final structure similar to the consonant-vowel relations in other languages. Each base syllable consists of either an initial followed by a final or a single final. Here initial is the initial consonant part of a syllable and

final is the vowel part including an optional medial or a nasal ending. In Mandarin Chinese, there are 22 initials (including a null-initial) and 38 finals as shown in the table (Hon, 1994; Wu, 1989).

2.1.3 Mandarin

Spoken Chinese exists in the form of different dialects. For example, Cantonese is spoken in Hong Kong and southern China. Mandarin is the standard spoken language of Chinese. Mandarin (Putonghua) is defined as "the common language in China, based on the northern dialects, with the Peking phonological system as its norm of pronunciation." (NLRM, 1955). In this thesis, in the context of speech, we use Chinese to mean Mandarin.

2.2 Chinese Prosody

The researches in Chinese prosody provide us a picture of Chinese prosody. Prosody of Chinese is unique in several ways. We briefly introduce the following: tone, intonation, and rhythm.

2.2.1 Tone

Chinese is a tonal language, in which each syllable (or Chinese character) carries a tone. Tone helps to express meanings in Chinese. The tone can be perceptually identified by human or observed from pitch analysis result. When a syllable is pronounced in isolation, its pitch contour is quite stable. Pitch contour of each tone is regular, except for tone 5, traditionally termed neutral tone, which is not considered as a formal tone. The pitch contour of base syllable "ma" is shown in Figure 2.2. (Xu, 1997). From the figure, we see that each tone has its shape.

However, when pronounced in a context, the pitch contours of tone undergo substantial variations, which usually depend on the contextual tones and sentence intonation. There are anticipatory effect and carry-over effect in Chinese tones (Xu, 1997). Pitch contour will change to have a smooth transition between itself and the contour of its previous syllable or the succeeding syllable. These effects exist between

syllables, even if the syllables do not form a word, as long as there is no pause between them.

It is well known that a third tone will be changed to the second tone when it is followed by another third tone. For example, the original pronunciation of "雨伞" (umbrella) is "yu3 san3". However, it is usually read as "yu2 san3".

It is possible for a prosodically weak syllable to be toneless, i.e. neutral tone (Tone 5). In extreme cases, a tone may be realized with a shape opposite to the lexical specification (Shih et al, 2000). The pitch contour of the neutral tone syllable is conditioned primarily by the tone of the preceding syllable, although other factors such as the following syllable also play a role.

From the above facts, we understand that pitch contour of a tone is heavily affected by the surrounding syllables.

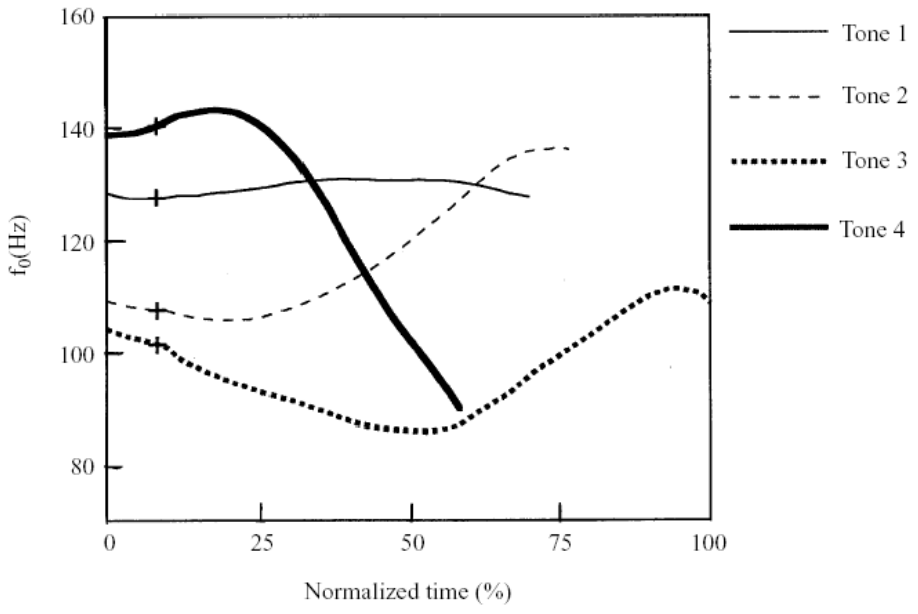


Figure 2.2 Tones and pitch tracks of base syllable "ma" (Xu, 1997)

2.2.2 Intonation Theory of Chinese

Unlike English and other non-tonal languages, in which the F0 contour is principally determined by intonation pattern alone, F0 in Mandarin Chinese also reflects lexicon tone for the component words. When syllables are stressed, their tonal shapes are fully realized, while weakly stressed syllables are usually overridden by sentence intonation. (Liao, 1994)

There are three different models previously proposed to describe intonation of Mandarin Chinese (Jin, 1996). (1) The pitch range theory (Gärding, 1987) claims that Mandarin intonation is a combination of different pitch range values determined by the sentence. Tones are just local pitch perturbations within the given ranges. (2) The pitch contour theory (Chao, 1968) claims that Mandarin intonation is characterized by contrasting contour shapes. These contours provide global rises or falls onto which the local word tone contours are superimposed. (3) The register theory (He and Jin, 1992) claims that Mandarin intonation contours are exhibited on different registers according to grammar and the speaker's attitude.

From these theories, we understand that Chinese intonation has a global shape for the whole intonation and local shape for tones. The global shape and local shape interact with each other.

2.2.3 Rhythm

One example of rhythm in Chinese is the existence of prosodic word. Linguistic research on Chinese prosody (Feng, 1997) found that the prosodic word in Chinese includes at least one foot, which is the smallest free-used prosody unit in prosody morphology. A standard foot in Chinese is bi-syllablic. Tri-syllable foot (super foot) and monosyllable foot (degenerate foot) are variations of standard foot. Super foot and degenerate foot are realized under certain conditions. When there is a single syllable around a standard foot, the syllable will be attached to the neighboring foot to form a super foot (Shih, 1986). Degenerate foot occurs in the case that a monosyllabic word constitutes an independent intonation group (Feng, 1997).

This indicates that sometimes, a monosyllable word will be attached to its neighboring words to form bigger prosodic unit. However, sometimes, a monosyllable word can stay alone in speech.

2.3 Classification and Regression Tree (CART)

Many parts of this research use the decision tree approach. CART approach (Breiman et al, 1984) is used as the main learning approach to construct decision trees. A decision tree is a tree structure that represents a classification system or predictive model. The tree is structured as a sequence of simple questions, and the answers to these questions trace a path down the tree. The leaf node reached determines the classification or prediction made by the model. A decision tree in general is tree-structured classifier that attempts to infer an unknown variable from an observed feature vector. The CART approach has some advantages:

- The sequence of the questions is automatically determined from training data.
- During the construction process, the important factors are automatically selected as question, while irrelevant factors are ignored.
- The relative importance of the feature can be examined from the tree that is constructed from the training data.
- The size of the tree can be easily scaled according to different needs.

2.3.1 Classification Tree or Regression Tree

Classification tree and regression tree are both types of decision tree where predictions are made based on questions about feature vectors. Classification trees assign a class based on the observed features. Regression tree are used to predict a continuous-valued variable. Both classification tree and regression trees are used in different parts of this research.

Many algorithms for constructing decision tree have been proposed, such as C4.5 by Quinlan (1993), CART by Friedman et al (1984). Wagon tool in Festival (Black et

al, 1998) is used as our main tool in the work. Apart from the predicted value, the leaf node for regression tree and classification tree can provide more parameters. For example, a regression tree can provide a standard deviation of the predicted value, while a classification tree is able to provide the probability distribution of each class in the node.

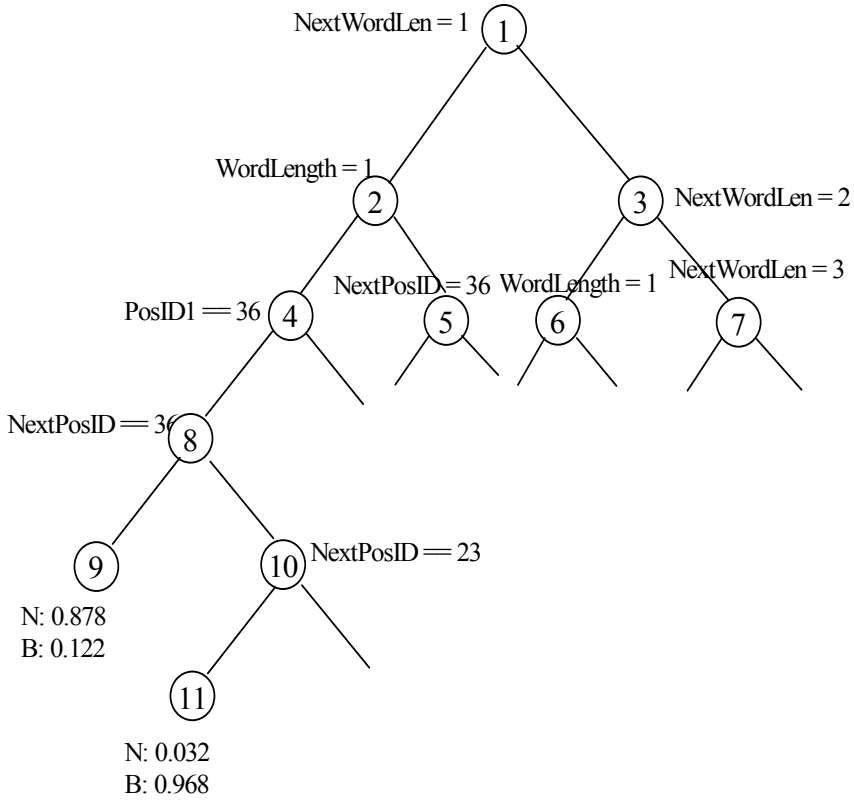


Figure 2.3 Example of classification tree
(Answer "yes" to left, "no" to right child)

Figure 2.3 gives an example of a classification tree, in which each node has a question based on the features of a feature vector. If an answer of a question is yes, the prediction goes to the left branch of the subtree. It goes to the right if the answer is no. Leaf nodes give the predicted values. For the feature values (NextwordLen = 1, WordLen = 1, PosID1 = 36, NextPosID = 3, and NetPosID = 3) of feature vector, the features trace a path from node 1, via node 2, node 4, node 8, and end at node 9. The

predicted value (at node 9) produces the result class N (or a probability of 0.878 of being class N, and a probability of 0.122 of being class B).

Figure 2.4 gives an example of a regression tree. For feature values (EndOfPW = 0, InitialID = 2, FinalID = 27, PosID = 14), the prediction traces a path from node 1, via node 2, node 4, and node 8, down to node 9. The predicted value is 0.126 with a predicted standard deviation of 0.023.

Generally, a constructed classification tree or regression tree works like a function

$$y = F(X) \tag{2.1}$$

where *F(X)* is the function that transforms the feature vector *X* into a value *y*. For regression tree, *y* is a continuous value, while for classification tree, *y* is an integer indicating a category.

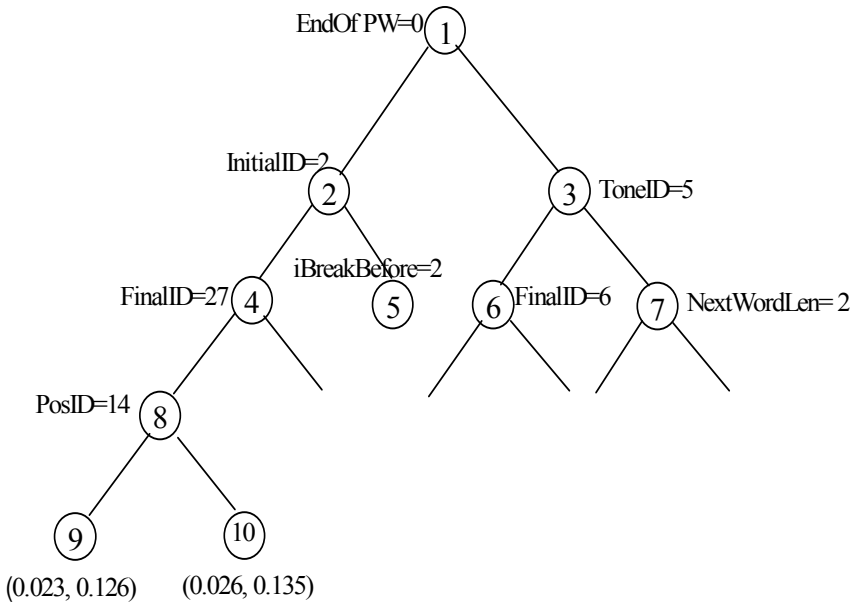


Figure 2.4 Example of regressin tree (Answer “yes” to left, “no” to right child)

2.3.2 Splitting Criteria

A tree grows by splitting the training data set. CART uses binary splits that divide each parent node into exactly two child nodes by posing questions with yes/no answers at each decision node. CART searches for questions that split nodes into relatively homogenous child nodes. As the tree evolves, the nodes become increasingly more homogenous. An impurity function is used in the classification trees to evaluate the goodness of the splits. A node's impurity function should be largest when it contains an equal mix of classes, and smallest when the node contains only one class. The different splits possible at a node are judged by calculating the decrease of the impurity of the whole tree. Each selected split tries to make the maximal decrease in impurity.

The decrease of impurity can be defined as:

$$\Delta i(t,s) = i(t) - P_R i(t_R) - P_L i(t_L) \qquad (2.2)$$

where the split s of node t puts a proportion P_R of data to the right child t_R and P_L to the left child t_L, and $i(t)$ is impurity function for node t.

There are different options to define impurity functions. Four types of impurity functions are commonly used in classification tree (Brieman, 1984). In Wagon, to make sure the splits of data will not make too small partitions, the program uses the following definitions as impurity functions:

(1) Regression tree: For sample sets with continuous predictees, impurity function $i(t)$ is defined as:

$$i(t) = v(t)N_t \qquad (2.3)$$

where $v(t)$ is the variance of the sample points in the node, N_t is the number of sample points in the node. The variance alone overly favors very small sample sets. Multiplying each part with the number of samples will encourage larger partitions, which will lead to more balanced decision trees in general.

(2) Classification tree: For sample sets with discrete predictees, impurity function *i(t)* is defined as:

$$i(t) = e(t)N_t \tag{2.4}$$

where $e(t)$ is the entropy of the sample points in the node, N_t is the number of sample points in the node. Again, the number of sample points is used so that small sample set is favored.

2.3.3 Building Better Tree

In the training process of a decision tree, a tree can be split small enough to make the tree work well for the training samples. However, the constructed tree is not necessarily good for data outside the training data. It is more desirable to build a classification/regression tree that will work well for new unseen samples. Some of the ways to make a better tree are as follows:

1. **Controlling the size of node.** The method is to build a full tree but make sure that there are enough samples in each node. An absolute minimal size for a tree node can be assigned. Alternatively, the minimal size can be a percentage number of the complete training data. The splitting of a tree stops when the splitting forms a node with size smaller than a stop value.

2. **Holding out data for pruning.** Another way is to hold out some of the training data for pruning. A tree with a small node size is first built and then pruned to where it best matches the held-out data.

3. **Stepwise training.** A good technique in Wagon is to build trees in a stepwise manner. In this case, instead of considering all features in building the best tree, it builds trees looking for which individual feature best increases the accuracy of the built tree on the provided testing data. Normally, a splitting process is used to look for the best question over all features. This technique first builds a tree using each individual feature that could lead to the best tree. Features are added one by one. This process continues until no more features

are added to the accuracy or some stopping criteria (e.g. size of node) is reached.

4. **Cross validation.** Cross validation is widely used in machine learning. By dividing the whole data set into different partitions, in each test, one partition will be reserved for testing, while the others work as the training data. This approach can generate a good result without bias.

2.4 Formulas

2.4.1 Mutual Information

The mutual information of two random variables X and Y with a joint probability mass function *P(x,y)* and marginal probability mass functions *p(x)* and *p(y)* is defined as (Cover et al, 1991):

$$I(X,Y) = \sum_{x_i \in X, y_j \in Y} p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \qquad (2.5)$$

Mutual information can be used to measure mutual dependency between discrete variables.

2.4.2 Pearson Product Moment Correlation Coefficient

Correlation coefficient is usually used to measure the dependency between continuous variables. Correlation coefficient between variable X and Y is defined as:

$$r = \frac{\sum_{x_i \in X, y_j \in Y} (x_i - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_{x_i \in X} (x_i - \bar{x})^2 \sum_{y_j \in Y} (y_j - \bar{y})^2}} \qquad (2.6)$$

Chapter 3 Speech Corpus Construction

In this chapter, the process of constructing the corpus is described. The distribution of speech units in Chinese language is investigated. The corpus is evaluated by the coverage of speech phenomena.

3.1 Speech Corpus Construction and Processing

Early systems used some rules to generate prosody parameters. Since too many factors affect prosody parameters and the factors interact with each other, it is difficult to use rules to cover all the factors. It is wise to use corpus-based approach, in which rules for the parameters can be derived by learning by analyzing speech corpus.

3.1.1 Consideration of Number of Speakers

The corpus in this work is produced by a professional female speaker. The reason to use corpus of only one speaker is as the following:

(1) The speech corpus will be used as unit inventory. A TTS system requires that all the speech units in the synthetic sentence come from the same speaker. Multiple voices are not usually used because we want to generate understandable and pleasant speech for general use. It is strange to have multiple voices in one utterance.

(2) The speech corpus is used for prosody training. The speaker for this corpus is a professional broadcast speaker. Her speaking style is considered as a good example for general listeners. As we want to generate speech with good prosody, we use the prosody contained in the corpus as our standard prosody style. Using multiple voices does not help to achieve this goal.

(3) Unlike speech recognition application, where it is desired to accept different styles of speech, a text-to-speech system is to generate a specific voice of speech. Therefore, a text-to-speech system is a speaker dependent system.

(4) This work uses corpus of one speaker. However, the approaches used in this work are not limited to this corpus. The approaches can also be used for new corpus. When we need to generate multiple voices in text-to-speech, we need to generate multiple corpora of single voice.

(5) Multiple-speaker speech corpus is useful when we want to investigate the general nature of speech of this language. However, this is not the aim of the work.

Due to the same reason as speaker, the corpus does not intend to cover different localities, different genders, etc. The female speaker of this corpus carries a Beijing style Mandarin accent, which is accepted as a standard spoken language in China and other parts of the world.

3.1.2 Speech Data

Since each Chinese character is a syllable, it is quite natural to use syllable as our analysis unit. In order to find the relationship between the text and the target prosody in the synthesis process, the speech need to be labeled with prosody data and the text should be analyzed and converted into a well-formed format.

The construction of the corpus mainly consists of the following steps:

Script design: In this research, the script for the speech recording is carefully selected using a greedy algorithm (Sproat 1997), which tries to cover as many pronunciation combinations as possible. The script is selected sentence by sentence from a huge text corpus. The huge text corpus consists of around 400M Chinese characters. The content of the text comes from many sources. Most of them are from Chinese web pages. The content of the text covers different styles of articles, including news, review, science, story, and so on. Finally, a large collection of sentences is selected. The average length of sentence selected is about 11. The selection process is not a part of this work. In this work, we use part of the collection as our corpus, which consists of around 3,600 sentences. The nature of our selected sentences will be discussed later in this chapter.

Recording: A professional female broadcasting announcer reads all the text in a neutral manner with normal speed. The recording is conducted in a studio designed for speech recording. The speech is recorded using a digital audiotape recorder at a sampling rate of 44,000 samples/second and a resolution of 16 bits. The recorded speech is then segmented into speech utterances of sentence and is stored in waveform files. If a mistake is made, the sentence is recorded again. A glottal wave device is used in the recording process. This device is attached to the neck of the speaker in order to record the glottal wave, which is the source of fundamental frequency. The glottal wave will be used for accurate calculation of fundamental frequency values.

Segmentation: Segmentation is to label continuous speech into small unit that is easy to manipulate. In this work, we use HMM-based recognition techniques to perform automatic segmentation. The segmentation is achieved by force aligning speech with text.

Manual verification: The segmented speech is then checked by human to remove any mistakes during automatic labeling process and to find any incorrectly read units. The sentences found with mistakes are read, segmented, and labeled again.

Pitch value calculation: One of the most important prosody elements is pitch contour. As we have recorded the glottal waveform, the glottal waveform is used for pitch calculation. This pitch extraction work is done using pitch extraction tool from Festival speech synthesis package.

3.1.3 Text Data

Text Normalization: The script text is first cleaned. The numbers are changed to corresponding Chinese characters. The symbols are removed. Therefore, the text is changed to pure Chinese text.

Word Segmentation: The word segmentation used HMM-based segmentation approach, which is trained on 6 months of People's Daily of PKU Tagged Corpus (Yu et al, 2002). A dictionary of around 60,000 words is used.

POS Tagging: An HMM-based tagging program, which is trained using PKU (Peking University) Tagged Corpus and PKU tag set, is used for POS tagging. The tag set is as shown in Appendix A.

Text to pronunciation conversion: A conversion program is used to convert the text into Chinese pinyin transcriptions. To make sure there is no error, the converted pronunciations are manually checked.

Prosodic Breaks: Prosodic breaks are also labeled in text data. In our research, we label the breaks manually. The break types we defined include: prosodic word break, minor phrase break, and major phrase break. The breaks are labeled by one person first and then checked by two other persons. One example of the labeled breaks is shown in Figure 3.1, in which space marks prosodic word, "|" marks minor phrase, and "||" marks major phrase.

想着 要靠 卖画 为生的 画家| 固然不少||，但是| 却有着| 各自 不同的 难处||。

Figure 3.1 Example of Chinese prosodic structure

3.1.4 Data Attributes

The final data is a collection of information that represents text and speech with a multi-tier structure. The data can be described as shown in Table 3.1.

Figure 3.2 shows an example of the tiers for speech data. Waveform, F0 contour, and syllable label tiers are shown respectively. Example of tiers for text data is shown in Table 3.2. In the table, space marks prosodic word break, "|" marks minor phrase break, and "||" marks major phrase break. The speech data and text data are aligned syllable by syllable.

Category	Data tier	Description
Text	Normalized text	Pure Chinese characters with punctuation marks
	Word segmented text	Words are segmented
	Pinyin	Corresponding pinyin of Chinese characters
	POS	POS types of each word
	Prosodic break	Prosodic word and prosodic phrase breaks
Speech	Speech wave data	Speech data
	Speech label	Labels indicating the start and end point of each syllable
	Pitch contour	Pitch contour of speech. The pitch value is given every 0.001 second. Unvoiced part is given a pitch value of 0.

Table 3.1 Data tiers of the corpus

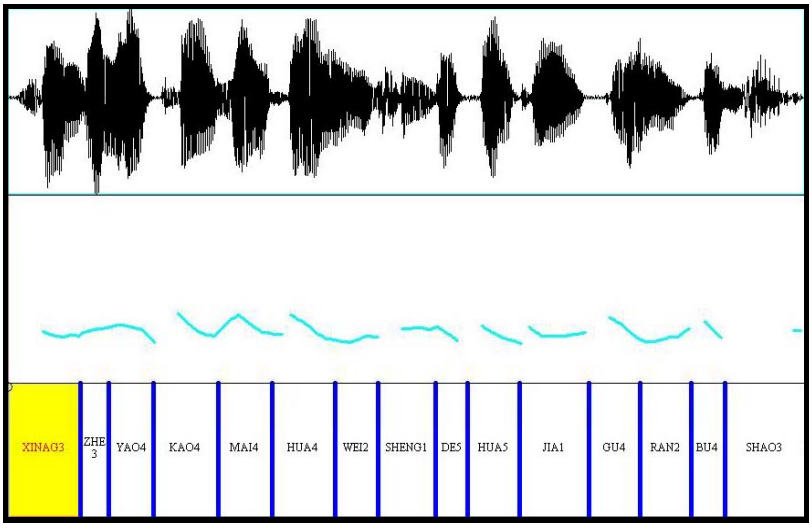


Figure 3.2 Example of speech tiers in the corpus
(waveform, F0 contour and syllable labels)

Tiers	Example
Normalized text	想着要靠卖画为生的画家固然不少
Word segmented text	想 着 要 靠 卖画 为生 的 画家 固然 不少
Pinyin	xiang3 zhe5 yao4 kao4 mai4 hua4 wei2 sheng1 de5 hua1 jia1 gu4 ran2 bu4 shao3
POS (Aligned with words)	v u v v v n v u n d d a
Prosodic break	想着 要靠 卖画 为生的 画家| 固然不少||

Table 3.2 Example of text tiers in corpus

3.2 Phonetic Statistics of Chinese

Both prosody training and unit selection need a corpus that has a good coverage of basic speech units and combinations of speech units. Because a unit is usually affected by its context units, it is desirable for a corpus to have a full coverage of context dependent units. In this section, we investigate this possibility by looking at the distributions of speech unit in Chinese language.

We use a text corpus that consists of 6 months of texts from the People's Daily (a Chinese newspaper), which was word segmented and POS tagged by Peking University (Yu et al, 2002), as real world corpus for statistics. The corpus consists of about 11.4M Chinese characters.

The reasons why we choose People's Daily are as the following:

- The articles in the newspaper use formal Chinese languages, which are suitable for readers from a wide range of backgrounds.
- There is a wide coverage of different genres, such as general news, views, economics, education, social science, etc.
- The corpus was well word-segmented, tagged, and checked by Peking University. So the accuracy of the corpus is guaranteed.

- This corpus is publicly available. Anyone can easily verify some of the results obtained in this work.

The text is first transformed into pinyin format. Statistics is done based on pinyin transcription. Counting is done sentence by sentence. Sentence start and sentence end are regarded as a special pinyin (e.g. represented as #).

Using P_i, I_i and F_i to represent the pinyin, initial and final of ith syllable respectively, we considered the following basic combinations of units in our statistics:

- Context independent unit (Unigram): $<P_i>$
- Context dependent unit (Trigram): $< P_{i-1}, P_i, P_{i+1}>$, $< F_{i-1}, P_i, I_{i+1}>$

Here, a unit means the syllable with pinyin and tone. Context independent unit means, unit itself is considered when counting the units. Context dependent unit means, the previous and next units of the current unit are also considered in counting.

In the following sections, we want to know how many of the most frequent units can have a good coverage of the real world text. The accumulative coverage of the first n frequently occurred units $c(n)$ is calculated by

$$c(n) = \sum_{i=1}^{n} f_i \Big/ \sum_{i=1}^{N} f_i \qquad (3.1)$$

where f_i means the frequency of the ith unit, N is the total number of units in the corpus. The units are sorted in descending order of frequency.

3.2.1 Context Independent Unit

Figure 3.3 shows the coverage of syllables in the corpus. X-axis is the number of syllables sorted by descending order of frequency. Y-axis is the accumulative coverage percentage. Totally, there are 1,373 distinct syllables in the corpus. The figure shows that around 400 most frequent syllables occupy around 90% of all the occurrences in the text corpus. This result shows that the distribution of syllables is quite unbalanced in the corpus.

3.2.2 Context Dependent Unit

In voice production process, neighboring sounds interact each other. This leads to phonetic coarticulation. Therefore, a voice inventory of rich contextual consideration is crucial to a quality TTS system. However, context of a unit can be considered at different levels. For example, we can consider the whole syllable of previous unit as left side context of the current unit. Alternatively, we can use final part of the previous unit as left side context of the current unit.

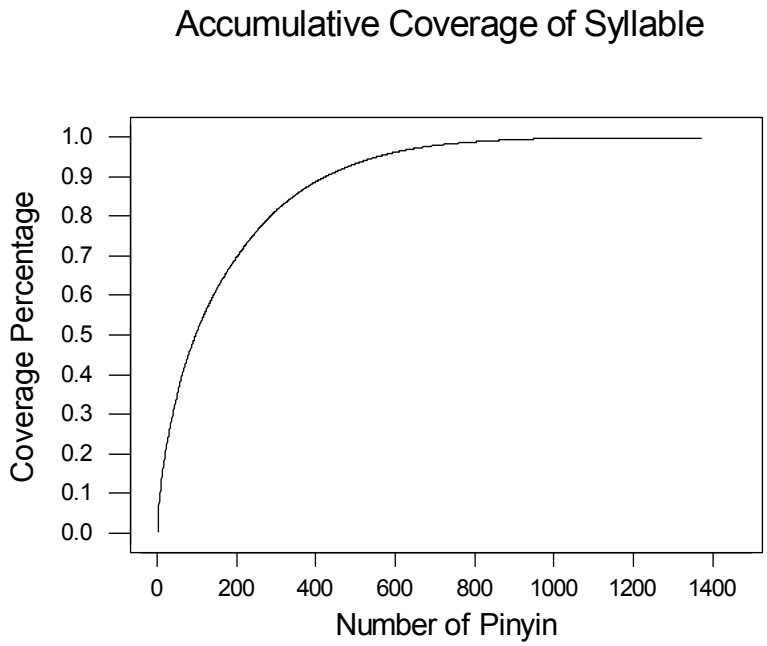


Figure 3.3 Accumulative coverage of syllables in text corpus.

We first consider the context of a syllable by looking at the pinyin (with tone) of previous and next syllables. In this consideration, the trigrams of unit ($< P_{i-1}, P_i, P_{i+1}>$) should be counted. Figure 3.4 shows the relationship between the number of units and the accumulative coverage of the units in the corpus. X-axis is the number of trigrams sorted by percentage in descending order. From the figure, we see that to cover 80% of $< P_{i-1}, P_i, P_{i+1}>$ occurrences in the corpus, we have to cover around 69×10^4 distinct units; to cover 90%, we have to cover 150×10^4 distinct units. It seems impossible to build a speech corpus to have such coverage.

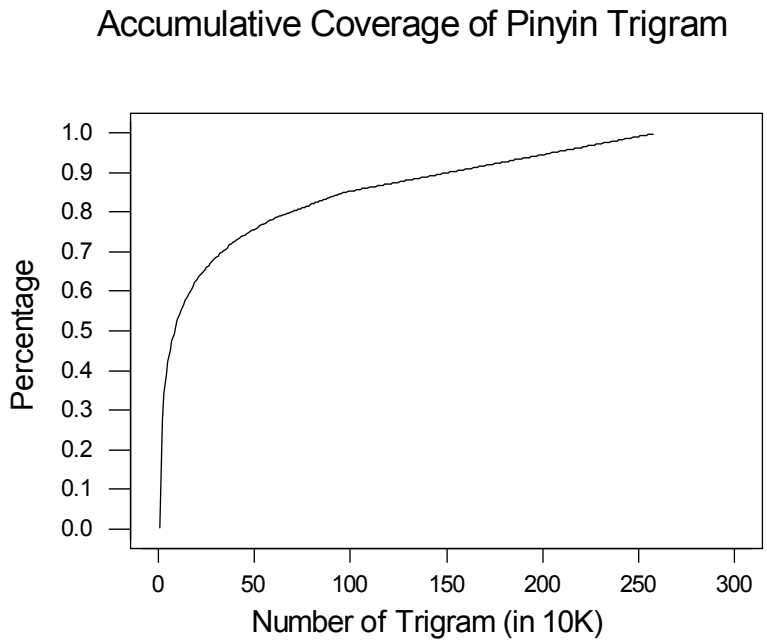


Figure 3.4 Accumulative coverage of pinyin trigram

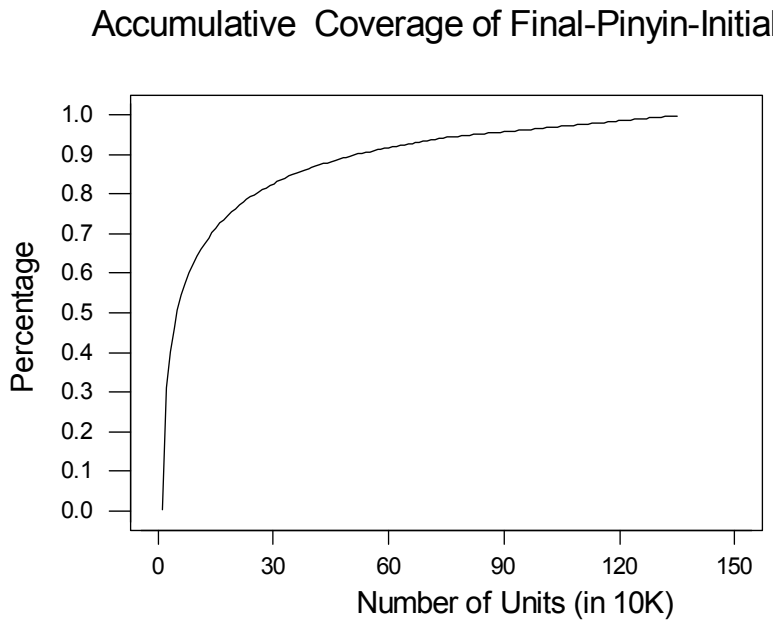


Figure 3.5 Accumulative coverage of syllable with context considered

We can narrow the scope of the context. Instead of considering full syllables of previous or next syllable, we can consider to use half of a syllable. If we only use the final and tone of the previous syllable, and initial and tone of the next syllable for context consideration, we have a coverage curve as shown in Figure 3.5. From the figure, we can see that to cover 80% of $< F_{i-1}, P_i, I_{i+1} >$ occurrences in the corpus, we have to cover around 26×10^4 distinct units; to cover 90%, we have to cover 52×10^4 distinct units. It seems that building a speech corpus with such coverage is not realistic either.

3.2.3 Grouping Context Units by Initial and Final

We can further reduce the number of context elements by clustering initial and final into initial and final class. For example, finals of the previous syllable "A" and "IA" can be grouped into one class because their coarticulation effects are similar.

Each syllable has a neighbor at its left side. There are 38 finals in Chinese. Therefore, the right edge of the previous syllable has 38 choices. In this work, we classify the 38 finals into 10 classes as shown in Table 3.3.

Class	Right edge of a syllable
R1	A IA UA
R2	AI EI I UI UAI IZ IZH
R3	AN EN IAN IN UAN VN VAN UN
R4	ANG ENG IANG ING IONG UANG ONG
R5	AO IAO O UO
R6	E
R7	ER
R8	IE VE
R9	IU OU U
R10	V

Table 3.3 Class of right edge (final) of syllable

For right neighbor of a syllable, there are 22 choices of initial. However, when the right syllable has a null initial, the final is the actual left side of the syllable. There are 11 finals that possibly follow a null initial. Therefore, there are 32 choices for left edge of a syllable. We classify left edges into 11 categories according to their production manners as shown in Table 3.4.

Based on the initial and final class, the number of context dependent unit can be calculated again. Tone of previous and next syllable can also be ignored if we want to further reduce the context consideration.

Class	Left edge of a syllable
L1	A AI AN ANG AO
L2	B D G K P T
L3	C CH F J Z ZH X S SH Q H
L4	E EN ER OU
L5	EI
L6	L
L7	M
L8	N
L9	W
L10	R
L11	Y V

Table 3.4 Class of left edge (initial or final for null-initial syllable) of syllable

3.2.4 Considering Loose Coarticulation

Considering context is to keep the coarticulation effects. However, there are different levels of coarticulation. When two units are succeeding units in an utterance, the coarticulation degree is determined according to pronunciation of the second unit (Wu et al. 2001). We define the following coarticulation types as follows:

- Loose coarticulation, when the initial of the second syllable is unvoiced.
- Intermediate coarticulation, if the initial of the second syllable is voiced.
- Tight coarticulation, if the second syllable has a null initial.

The actual initials are listed in Table 3.5.

Coarticulation type	Initials of the second syllable
Loose coarticulation	B C CH D F G H J K P Q S SH T X Z ZH
Intermediate coarticulation	L M N R
Tight coarticulation	NULL-INITIAL

Table 3.5 Classification of initials for tightness of connection.

Realizing the differences in coarticulation degree, we can further group units, which have loose coarticulation with its context. For right context, the initial of next syllable, L2 and L3 can be combined to one category because all of them belong to loose coarticulation. For the left context (final of the previous syllable), when the initial of the current unit belongs to loose coarticulation, the left context can be ignored because we can assume that this unit is not affected by its left context. Therefore, the number of unit and distribution of units can be calculated again based on the reduced context dependent unit set.

3.2.5 Unit Distribution for Different Context Considerations

We compare the coverage of context dependent units in Table 3.6 by different context considerations. In the table, we can see that if full pinyin (with tone) is used for consideration of the previous and next syllables, there are 2.57M different combinations in the corpus. If we only consider the previous final and next initial and tones on both sides, the total number of different combinations is reduced to 1.34M. Further grouping initial and final into class and considering tones on both sides, the total number is reduced to 481,590. Ignoring tones on both sides, there are 80,378 different context dependent units. If we further consider context with loose coarticulation has no great effect for the current syllable, the number of context dependent units is 26,972.

When we design a speech corpus for TTS, we usually have to construct a corpus of natural utterances instead of single units, i.e. each unit cannot be recorded in isolation. Rather, units should stay in carrier sentences in order to maintain

naturalness. The method inevitably keeps many redundant copies for many units. Therefore, to cover a specific number of units, the number of units in the final corpus could be many times larger than the number of units intended to cover. For example, to cover the 80,378 units, the size of the corpus should be much larger than 80,387. Realizing the difficulty in covering all possible units in a corpus, a corpus should be built with a fair coverage for a reasonable level of context consideration rather than a full coverage of possible variants of unit.

Previous Unit	Next Unit	10%	20%	50%	90%	100%
Initial Final Tone	Initial Final Tone	946	4,437	73,430	1.49M	2.57M
Final Tone	Initial Tone	841	3,474	37,996	500,693	1.34M
Final Class Tone	Initial Class Tone	464	1,608	12,052	118,284	481,590
Final Class	Initial Class	89	290	1,847	13,966	80,378
Final Class, Loose coarticulation collapsed	Initial Class, Loose coarticulation collapsed	11	36	253	2,770	26,973

Table 3.6 Number of units for coverage of context dependent units

3.3 Corpus Evaluation

Corpus is very important for this research. If a corpus does not reflect the language well, the result based on the corpus will be unreliable. As the corpus is used for both prosody training and unit selection, a good design of the corpus script should meet the following criteria:

- The text used for recording the speech corpus is a true reflection of the general text corpus.
- The speech corpus used for unit selection has enough pronunciation coverage of Chinese language.
- The corpus should be sufficiently large. This allows that it has enough occurrences for individual features. For example, for training of prosody, we need each tone to have enough occurrences in corpus.

3.3.1 Word Frequency

To use a statistical approach, the corpus we used should be a good reflection of real world data. We use the PKU People's Daily Corpus as our reference corpus to represent the real world text. Because the models for prediction of prosodic structure will be based on the words in corpus, we will compare the relative frequency of words with that of words in reference corpus.

We calculate the relative frequency of common word in the two corpora. A correlation value of 0.93 is achieved between the frequency values. From the correlation value, we see that the content of the speech corpus is positively correlated with the reference corpus.

3.3.2 Syllable Coverage

We investigate how the speech corpus covers the real world units. There are 1373 distinct syllables in PKU corpus. Our corpus consists of 1261 distinct syllables, which cover 96.58% of all distinct syllables (1326/1373) and cover 99.88% of occurrences of syllables in PKU corpus. We see from the data that the corpus has a good coverage of context independent units the language. This means, in a real TTS process, most of syllable can be found in the speech corpus.

Previous syllable	Next Syllable	Percentage covered
Final class, tone	Initial class, tone	33.1 %
Final class	Initial class	76.8%
Final Class, Loose coarticulation collapsed	Initial Class, Loose coarticulation collapsed	90.4%

Table 3.7 Coverage of context dependent units of the corpus

We also consider the coverage of context dependent units. The result is shown in Table 3.7. The constructed corpus covers 76.8% of the context dependent units if context is considered using final class for previous syllable and initial class of next

syllables, and covers 90.4% of context dependent units if loose coarticulation is considered as not having effects on this current unit. Therefore, in our consideration, at least 90.4% of the units can be synthesized seamlessly almost without problem in coarticulation.

3.3.3 Statistics

We stated that the corpus should be sufficiently larger so that there are enough occurrences for individual features. However, this does not mean that we need that every feature should have many occurrences. Because language is unbalanced itself, overlooking very rare cases does not damage the general accuracy of this work.

In this part, we will give some details on the coverage of some basic units in this corpus. From the number of occurrences, we will have a better understanding of the nature of the corpus.

The corpus includes 38,713 Chinese characters in 3,609 clauses or sentences. The average length of a sentence is 10.7 (38,713/3,609) characters. The numbers of units are as shown in Table 3.8.

Unit	Number
Characters	38,713
Words	27,293
Prosodic words	17,040
Minor phrases	6,341
Major phrases	3,682
Sentences	3,609

Table 3.8 Number of text units and prosodic units in the corpus

Word Length	Number of Occurrence	Unique word
1	17,547	2,409
2	8,979	4,826
3	517	381
4	250	223
Total	27,293	7,839

Table 3.9 Length distribution of words in the corpus

Among the 27,293 words, there are 7,839 unique words. The numbers of words in different lengths are shown in Table 3.9.

POS	Frequency	POS	Frequency	POS	Frequency
Ag	158	e	7	nz	33
Bg	2	f	516	o	15
Dg	25	g	417	p	1,292
Ng	1,007	h	4	q	763
Rg	25	i	148	r	1,287
Tg	34	j	288	s	110
Vg	309	k	68	t	156
A	1,232	l	60	u	1,661
Ad	129	m	1,338	v	6,021
An	65	n	5,355	vd	15
B	123	nr	977	vn	562
C	822	ns	241	y	122
D	1,843	nt	10	z	53

Table 3.10 Frequency of POS in corpus

Num of occurrences	Num of Syllable	Coverage of base syllables
>= 300	12	2.9%
>= 200	27	6.6%
>= 100	88	21.6%
>= 50	161	39.5%
>= 20	284	69.6%
>= 10	304	74.5%
>= 5	375	91.9%
>= 1	400	98.0%

Table 3.11 Occurrence distribution of toneless syllable in the corpus

Tone	Number of Occurrences
1	5,529
2	7,530
3	5,230
4	9,640
5	2,142

Table 3.12 Distribution of tones in the corpus

The number of words falling in each POS category is also counted as shown in Table 3.10. We see from the table that the distribution of POS is unbalanced.

We also counted the occurrences of pronunciations in the corpus. There are 977 distinct syllables, and 378 toneless syllables occurred in the corpus. Table 3.7 shows the number of occurrence of distinct toneless syllables in this corpus. We can see that 74.5% of the base syllables have at least 10 occurrences in the corpus.

The occurrences of tones are shown in Table 3.8. Tone 4 has the most occurrences in corpus, and tone 5 has the least occurrences.

3.3.4 Conclusion

From the discussion and data provided in previous parts of the section, we understand that the speech corpus is a close approximation to the real world text corpus. The speech corpus has a good coverage of context independent units and a fairly large coverage of context dependent units.

In brief, the constructed corpus is a reflection of real world text with a little bias to cover as many context dependent unit variants as possible.

3.4 Summary

In this Chapter, we have described the process of constructing the speech corpus. A study of context independent units and context dependent units has been conducted. We have understood that building a speech corpus with full coverage of context dependent units is not realistic. We have evaluated the corpus and found that the corpus used in this research has both a good coverage of context independent units and a large coverage of context dependent units.

In the evaluation of speech corpus, I use some approaches to reduce the number of context dependent units. This solution reduces the number of context dependent units significantly. It makes building small speech inventory for text-to-speech synthesis possible. It also provides solutions for building text-to-speech inventory of different scales.

Chapter 4 Prosodic Break Prediction

In this chapter, we discuss the prosodic break (or unit) prediction, especially the prediction of prosodic words. First, an introduction is given. Then the determinations of prosodic word and prosodic phrase are discussed. The problems are described; the prediction models are presented; and the experiments are conducted.

4.1 Introduction

4.1.1 Prosodic Break

When speaking, people tend to group words into small prosodic unit groups. This occurs not only in Chinese but also in other languages. Grouping words into phrases helps the speaker to speak more easily and the listeners to understand the sentence better. Punctuation marks are explicit symbols in a sentence to indicate breaks. However, more breaks will be inserted within a long sentence when we speak. The sentence is therefore broken into short phrases, which are called prosodic groups. Take an English sentence for example. The sentence “I went to the bookstore in order to buy a book.” can be read as “I went to the bookstore [break] in order to buy a book”, but it is unusual to read it as “I went to the bookstore in [break] order to buy a book”.

Prosodic phrase boundary can be identified by some pauses, pitch changes, or duration changes of boundary syllable in speech. In a TTS system, to realize all these effects in synthetic speech, phrase boundaries need to be determined first. Phrase boundary is realized by inserting pauses, changing the pitch contour, and lengthening duration of the boundary syllables, etc.

It is common that words representing a meaningful concept are grouped into a phrase. However, prosodic phrase does not always coincide with the phrase from a syntactic point of view.

Usually phrase boundaries are also referred to as prosodic breaks. Different levels of prosodic unit (or break type) can be defined. In English, the ToBI labeling system (Silverman, 1992) defined six types of break. In some researches, less break types were defined.

This work is to predict prosodic break for Chinese text-to-speech.

4.1.2 Review of Existing Approaches

Many approaches have been proposed for the determination of the prosodic breaks (or units) from text input. Typical approaches include rule-based and corpus-based ones.

Rule-based approaches were first used for locating phrase boundary. MITalk (Allen et al, 1987) parses text into noun phrases, verb phrases, and prepositional phrases. The phrases are defined by grammar rules. After obtaining the phrases and clauses, pauses are inserted to break up long sentences.

Liberman and Church (1992) proposed a very simple but efficient approach, which defines phrase by at least one function word followed by at least one content word. The parser first searches for the function words then searches for the content word for each function word. Break is placed before each function word that follows a content word. Despite its simplicity, it produces better results than the approach in MITalk. This is because, in English, boundaries are more likely between content words and function words, because most functional words are placed before the words they are related. Note that this is not necessarily true for other languages.

Bachenko and Fitzpatrick (1990) proposed another rule-based method, which transforms a given syntactic tree to a prosodic tree via several rules. It finds prosodic phrase breaks from boundary salience indices that are generated from the level of bracketing between words in a full syntactic parsing.

Typical corpus-based approaches include classification and regression tree (CART), neural networks, and hidden Markov models (HMM). Wang and Hischberg (1992) used CART for locating English phrase boundary in ToBI framework. They used POS, time-based and word-based distance, and syntactic information as features

for CART trees. Ostendorf and Veilleux (1994) have developed two automatically trained models for predicting prosodic phrase breaks, a decision tree model, and a hierarchical model. The models use text-based features, which includes punctuation, POS, and syntactic constituency. The decision tree approach is able to use text features within a large window of utterance and is able to take into account the break dependency using Markov relations between breaks. The hierarchical model represents prosodic phrasing of entire utterance as a nested hierarchy of constituent phrases. Decision tree was used to represent the lowest level constituent.

Fujio et al (1994, 1995) presented models for predicting major phrase boundary location and pause insertion for Japanese using a stochastic context-free grammar (SCFG) from an input word class sequence. These prediction models were made with similar idea, as both major phrase boundary location and pause insertion have similar characteristics. In these models, word attributes and left/right-branching probability parameters representing stochastic phrasing characteristics are used as input parameters of a feed-forward neural network for the prediction. To obtain the probabilities, first, major phrase characteristics and pause characteristics are learned through the SCFG training using the Inside-Outside algorithm. Then, the probabilities of each bracketing structure are computed using the SCFG. Experiments were carried out to confirm the effectiveness of these stochastic models for the prediction of major phrase boundary locations and pause locations. Accuracy of 85.2% for pause boundaries and 90.9 % for no-pause boundaries were achieved.

Taylor and Black (1998) proposed another method, which uses only POS information. The sentence is first converted into POS sequence. Then a Markov model is used to give the most likely sequence of breaks.

Sun et al (2001) used decision trees to estimate the probability of a word juncture type (break or non-break) given a finite length window of POS values, and used an n-gram model to choose the word juncture sequence. Trained on an 8,000 word database, the algorithm predicted breaks with F=77% and non-breaks with F=93%. (F is a combined parameter indicating precision and recall)

However, the above-mentioned approaches have one or more of the following limitations:

(1) Some of the approach are rule-based and language dependent. For example, Liberman and Church, (1992) and Bachenko and Fitzpatrick (1990) are rule-based and language dependent.

(2) Some of them require parsing of sentences, which is slow, inaccurate, and unsuitable for a practical TTS implementation. For example, in Fujio et al (1994, 1995), prediction of breaks is based on SCFG.

(3) Some approaches may have problem of data sparseness during calculation of probability. In Taylor and Black (1998), to have enough data to calculate probability, the models should be designed to have a limited span of only a few words.

(4) Prosody phenomena in Chinese are different from other languages. For example, people usually avoid using monosyllabic words in Chinese; hence, monosyllabic words are likely to be combined with their neighboring words in speaking. None of the above approaches is directly suitable for predicting Chinese prosodic word breaks.

The above reasons make the previous approaches for other languages unsuitable for Chinese language or the approaches need to be improved.

4.1.3 Review of Work for Chinese

Come back to Chinese, most early Chinese TTS systems inserted break after every word or used rules to determine some breaks. More recently, there were a few approaches for phrase determination. Chou et al (1998) proposed an approach to first form a lattice to include the possible phrase grouping, and then find the best path from the lattice according to the frequencies of POS grouping. Chen et al (2001) proposed an approach based on inductive learning algorithm and extension matrix theory. POS sequence and syntactic structure are used in the phrase model. POS type and length of constituents in terms of the number of characters and words are used as features for

prediction and a success rate of 93% achieved on 371 training sentences and 188 testing sentences.

Generally, some of the above researches for Chinese phrase break reported good prediction results. However, there are uncertainties or shortcomings in the above approaches.

(1) Prediction of prosodic phrase on the base of syntactic structure can achieve good prediction result. However, the accuracy of automatic syntactic structure parsing was not reported. Considering the errors occurred in syntactic analysis, the accuracy of the prosodic phrase break will be lower for a prediction from POS sequence.

(2) Most of them used small number of sentences in the experiments. Due to the large number of words and POS types and the richness of language phenomena in Chinese, they do not sufficiently show how the methods work well on larger corpora.

(3) Previous work only put efforts on phrase breaks. Prosodic word was regarded as a common prosodic phrase. However, prosodic word break has its own characteristics (rhythm requirement). It should be specially treated.

4.2 Determination of Prosodic Breaks

4.2.1 Chinese Prosodic Structure

Each Chinese character is pronounced as a syllable. Syllables form a word, and words are connected together to form a sentence. From the view of prosodic structure, prosodic units can be defined. Researches have found that there is a hierarchical prosodic structure for Chinese prosody, which constitutes the rhythm of Chinese speech (Shih 1986). There are following prosodic elements in Chinese speech:

- Prosodic word (PW): Prosodic word is the basic building block of rhythmic structure of sentence. A prosodic word usually consists of one, two, or three syllables. However, in most cases, it consists of two or three syllables. Prosodic word can be a single word, part of a word, or combination of words. For example, 4-syllable words may be taken as two disyllabic prosodic words.

- Prosodic phrase (PP): Prosodic phrase is a common rhythm unit in the production and perception of speech. It is usually a meaningful combination of prosodic words. Some researchers indicate that the span of the chunk is usually within nine syllables in Chinese (Cao, 2000).

- Intonation phrase (IP): Intonation phrase is a rhythmic group containing one or more prosodic phrases, and is usually identical to syntactically meaningful sentence. There is usually a long pause after an IP.

In this research, the following units are used in the corpus:

- Character: The smallest unit in text. Each character is a syllable in terms of pronunciation.

- Word: A word consists of one or more characters. It is a unit from the syntactic view.

- Prosodic word: In this research, short words are combined to form prosodic word. However, we do not split a long word into small prosodic words. Therefore, prosodic word is one single word or a combination of two or more words.

- Minor phrase: Small meaningful phrase in utterance. This is equivalent to the prosodic phrase that we have mentioned previously. A minor phrase usually consists of several prosodic words.

- Major phrase: Phrase with an obvious pause in an utterance. This is equivalent to the intonation phrase that we have mentioned before.

Each of the above unit marks a type of break. Therefore, the following breaks are defined: syllable break (SB), word break (WB), prosodic word break (PWB), minor phrase break (PPB), and major phrase break. (IPB) In the break sets above, each set is a subset of previous set. The later three breaks are termed as prosodic break in this work.

4.2.2 Issues of Prosodic Break in this Work

Although Chinese linguists have found that there is a hierarchical prosodic structure for Chinese (Shih 1986) for many years, the problem of prosodic structure was not well dealt with in Chinese TTS systems. Some systems attempt to use different types of break. However, the problem of prosodic word, which is a more basic prosodic unit, was not well studied. Some researchers are aware of the importance of grouping words into prosodic words. However, approaches for grouping the words were not well given. Usually, rules are defined for the prosodic word prediction. In this work, we will investigate the patterns of prosodic word and will propose corpus-based approach for prosodic word prediction.

As for phrase prediction, there were some attempts in using corpus-based approach to generate better prosodic break. Nevertheless, they rely on a parsing tree. Due to the difficulty in parsing a sentence and the existence of many ungrammatical sentences or phrases in real text, the approach of using parsing tree is not realistic in a real TTS system. In this work, we will propose a corpus-based approach to generate minor phrase breaks from word sequences.

(1) Corpus Issues

In this work, we use corpus of one speaker for prediction of prosodic word break and minor phrase break. The reasons to use one speaker can be explained as the following:

(1) This research is to predict breaks for an input text based on the corpus of one speaker. Therefore, the style of break placement is the same as that of the speaker. Because the speaker is speaking standard Mandarin with a normal style, the generated speech corpus is one representative of speech for people who use this language.

(2) We use the corpus of one speaker for our test. However, this corpus is much larger (3600 sentences) than corpora used in many of the research projects for prosodic break prediction, in which usually hundreds of sentences are used. Therefore, our corpus covers more phenomena of prosodic word breaks.

(3) The proposed approach is speaker independent because the prediction does not rely on speaker dependent features. It is possible that the speakers in different parts of world have different speaking styles. However, the phenomenon of prosodic word is one of the characteristics of Chinese language. The patterns of prosodic group may be different among different speakers from different parts of the world. However, as long as there exist such patterns, the approach proposed in this work will still work.

(2) Disagreements between Labelers

In the processing of the corpus, we labeled three types of breaks, which are prosodic word break, minor phrase break, and major phrase break. The former two takes a large part of all the breaks. To evaluate to what degree the labeling work can be reproducible, we ask three people to label 100 sentences and compare the results. We found that for prosodic word breaks, between labels of each two persons, there is around 96% overlap. However, for minor phrase breaks, this value is around 83%. Therefore, there is more agreement on prosodic word, but less agreement on minor prosodic phrase. The disagreements show that there is no clear definition for prosodic word and prosodic phrase. The result shows that there is more agreement of prosodic word than minor phrase among different people. This is due to different people have different breaking styles. They have more choices for minor phrase level breaks.

(3) Prosodic Unit and Prosodic Break

The work of this chapter can be considered as prediction of prosodic units. Since prosodic units are separated by prosodic breaks. Correctly predicting prosodic break also correctly predicts prosodic units (prosodic word and minor prosodic phrase in this work). Therefore, the work of this chapter can also be considered as prediction of prosodic break.

The task of prosodic break determination is first to combine short words to form prosodic word, and then combine prosodic words to form longer phrase, which may be uttered as a prosodic unit in speaking. The prosodic structure prediction work for Chinese includes the following parts: Prosodic word detection, minor phrase break prediction, and major phrase prediction.

The corpus we used here is a text transcription of our speech corpus. The three types of break are manually labeled. The final data are word sequences marked with different types of break. As most of the sentences in the corpus are short sentences (around 10 syllables), there is only one major prosodic phrase in a sentence. Therefore, in this research, we will only predict prosodic word breaks and minor phrase breaks. In prediction, we assume each sentence end is a major phrase break. Since the number of major phrase is small, this simplification will not greatly affect the accuracy in the prediction of prosodic structure.

Prosodic word break set is a subset of word break set. Therefore, the prediction of prosodic word is to determine which word break should also be marked as prosodic word break.

Similarly, minor phrase break set is a subset of prosodic word break set. Therefore, the prediction of minor phrase break is to determine which prosodic word break should be marked as minor phrase break.

(4) Approach Option

One of the ways to solve the problems of break determination is to use a single method to determine the different types of break, such as using decision tree approach. However, because the different break types are determined by different factors, it is better to predict different breaks separately. We understand that the phenomenon of prosodic word is a phenomenon that some monosyllabic words are attached to other words. However, minor and major phrases are grouping of meaningful words (Shih 1986).

In this work, we will first build model to predict prosodic word breaks. Then we build model to predict minor phrase breaks.

4.3 Prosodic Word Detection

The prosodic word detection problem is unlike the other prosodic break detection in that it is a local combination of words (demanding of rhythm) rather than a global

consideration of words (logically meaningful grouping). The detection of prosodic word is a process to find whether two words, which are usually monosyllable words or one of which is a monosyllable word, should be combined together to form a single prosodic unit.

4.3.1 Prosodic Word

In word segmentation, a string of Chinese characters is separated into Chinese words under the guidance of lexicon or rules. In our lexicon, only words of length from 1 to 4 are included. The word segmentation program looks for words that are included in the lexicon from the sentence. Some rules are used to find the words that are not in the lexicon. After word segmentation, a sentence is converted into a sequence of words.

However, the words are only defined from a syntax view. In real speech, people do not speak Chinese word by word as performing word segmentation. Instead, neighboring words are grouped together when speaking. Take the sentence "请把这本书给你哥哥" (Please pass this book to your brother) as an example. The result of word segmentation is like "请|把|这|本|书|给|你|哥哥|". However, in speech, the sentence is more likely to be read like "请把|这本书|给|你哥哥", in which "请把", "这本书" and "你哥哥" are uttered together respectively. Actually, each group of the words is a combination of some short words.

Prosodic word is a group of syllables that are uttered closely and continuously. Grouping of the prosodic word considers the meaning of word and rhythm of speech. In most cases, a prosodic word is a compound word or a meaningful unit. It is a concept of words based on prosody rather than syntax. Some prosodic words are actually phrases in a syntactic view. To distinguish from prosodic word, the words from word segmentation are called syntax words. Previous studies on Chinese prosody have shown that prosodic word is an important prosody unit in Chinese. (Qian et al, 2000).

The relationship between syntax words and prosodic words includes three types. (1) A prosodic word is a syntax word. (2) A prosodic word is combination of several

short syntax words. (3) A prosodic word is part of a long syntax word. In this research, 4-syllable words are also taken as one valid prosodic word because usually, the prosodic break is not obvious in many 4-syllable words.

The existence of prosodic words renders speech with rhythm. There are short breaks between prosodic words in sentence. The use of prosodic word in TTS includes the following (1) It gives correct breaks in the sentence. (2) It helps to make tone changes (tone sandhi) in the sentence. (3) It helps to improve the accuracy of prosody parameter prediction since prosody properties of boundary syllables are different from those of non-boundary ones.

In TTS, we need to find prosodic word from syntax word sequence. The problem of prosodic word detection can be considered as a problem of deciding whether there should be a prosodic word break or not between two syntax words, which is actually a classification problem. This work is to find rules to detect prosodic word breaks using corpus-based approach.

We considered the following parameters or constraints in prosodic word prediction: (1) What features of words are used in prediction? (2) How many categories of part-of-speech are suitable? (3) How many words should be specially dealt with in feature set? (4) How previous predicted break will affect the next prediction? (5) How the dependency between breaks will help to improve accuracy? (6) What parameters for CART are suitable for the experiments?

4.3.2 Patterns of Prosodic Words

We are dealing with how words can be combined to form prosodic word. It is not realistic to consider each word individually even if we have a very large corpus. Therefore, we have to use word group for generalization purpose. POS is a natural grouping method. To consider rhythm, we also need to consider the length of word (number of syllables in word). Therefore, we can think of the following features: (1) POS type of word. (2) Length of each word. The two features can be justified by looking at the patterns of prosodic word.

POS Patterns for prosodic word	Percentage
d+v v+v v+u m+q v+n n+n n+u v+p a+u a+n n+f v+Ng r+u r+v m+m v+r p+r m+n d+p n+Ng d+a p+n nr+nr adv+d+v m+m+q r+q r+n n+v v+a	50%
Other 1416 patterns	50%

Table 4.1 Prosodic word patterns in terms of POS

- **Patterns appeared in terms of part of speech.** Among 17040 prosodic words in our corpus, we found around 55% of them are single words, i.e. a syntax word is a prosodic word. Among the rest 45%, there are 1446 types of POS combination. Table 4.1 lists the patterns of prosodic word in terms of part of speech. The first 30 frequent POS patterns covers around 50% of all the POS combinations. We can see from the table that most patterns consist of two words. The mostly appeared POS types are noun, verb, adjective, and numeric words, which are represented by n, v, a, and m respectively.

Length patterns for prosodic word	Percentage
1+1	46.2%
2+1	15.6%
1+1+1	15.4%
1+2	7.0%
1+1+1+1	4.3%
2+2	2.9%
2+1+1	1.8%
1+1+2	1.4%
1+2+1	1.2%
1+1+1+1+1	0.8%
4+1	0.8%
3+1	0.7%
Other 20 patterns	1.8%

Table 4.2 Prosodic word patterns in terms of word length

- **Patterns appear in terms of length of word.** Table 4.2 lists the distributions of prosody word patterns in terms of length of word. In the table, "1+1" means the prosody word is composed of with two monosyllabic words. We can see that almost all patterns contain monosyllabic words, and the "1+1" pattern accounts for 46.2% of all occurrences.

Feature	Mutual Information
POS_{i-1}	0.0059
LEN_{i-1}	0.0322
POS_i	0.1031
LEN_i	0.0566
POS_{i+1}	0.1737
LEN_{i+1}	0.1361
POS_{i+2}	0.0064
LEN_{i+2}	0.0020

Table 4.3 Mutual information between break type and features

We use break type in the following discussions. Break type means a binary value to indicate whether, in a break position, there is a prosodic word break or not. For example, 1 means existence of prosodic word break, and 0 means non-existence of prosodic word break.

To determine the break type (1 for existence, 0 for not) between two words, we need to consider whether the two words can form a prosodic word. However, we cannot only look at the two words alone. We need to look at a wider range around the break. One choice is to choose a window of a few words (e.g. we choose four words) around the break (2 before and 2 after) for the prediction. This allows us to compare, among the three breaks between the four words, which one is the most possible prosodic word break.

Before building a model, we first examine some main features for the prediction. Mutual information is a measure to evaluate the dependence between events. We calculate the mutual information (For formula, refer to Section 2.4.1) between the break (break between word w_i and w_{i+1}) and POS types and length of words (POS_i

and LEN_i mean POS types and length of word $w_{i.}$). The calculated mutual information is shown in Table 4.3.

In the table, we can see that POS_i, POS_{i+1} , LEN_i, LEN_{i+1} have a larger mutual information value than other features. This shows that the two words immediately next to the word break have the largest effect on the prosodic word break types (existence or not).

4.3.3 Baseline Model

We will use CART approach for the prediction because CART has the following advantages: (1) It can incorporate different types of features, and there is no limit for the number of inputs. Therefore, we can add additional features to improve its performance; (2) it can automatically select the most important features for classification.

The data item for constructing a decision tree consists of a feature vector and an expected resultant value. Suppose we are to determine the break type (1 for existence and 0 for non-existence) between w_i and w_{i+1}. The feature vector includes the information of the four words (w_{i-1}. w_i, w_{i+1}, w_{i+2}) around the break. The basic features we used are:

- **POSs of w_{i-1}, w_i, w_{i+1} and w_{i+2} (POS_{i-1}, POS_i, POS_{i+1} and POS_{i+2})**: There are 35 POS types in our corpus. NULL is set as POS type if the word does not exist. (e. g., $w_{-1.}$)

- **Lengths of w_{i-1}. w_i, w_{i+1} and w_{i+2} (LEN_{i-1}, LEN_i, LEN_{i+1} and LEN_{i+2})**: The length of word is in the range from 1 to 4 because the lexicon has a maximum word length of four. If a new word has a length more than 4, the length feature is set to 4. Length is set to zero if the word does not exist. (e. g., w_{-1} is null.)

4.3.4 Grouping POS Categories

Due to the limited size of the corpus, there may be not enough training data for some POS types. To make the models more general for the POS types, we have to merge

some small POS types. In this work, we try the following approach to reduce the number of POS types and hence to improve generality.

- Simply merge POS types by the frequency. We combine most rarely appeared POS types in this work.

- Merge POS types according to its discriminating abilities. In the process of constructing the decision tree, when only one feature is used, the values are used one by one from root down to a single node. This will give us a sequence of POS types ordered by their discriminating ability. The POS types with less discriminating ability should be combined together because their low discriminating ability may be caused by insufficient occurrences.

4.3.5 Single Word Categories

On the other hand, some frequently used words may have their own characteristic different from the other words in the same POS category. Therefore, we need to form groups for some single words. In this work, we will define single word groups for the most frequently used words, and a separate group for the rest of the words. Therefore, we need additional features, which are:

- **Single word group type for w_{i-1}. w_i, w_{i+1} and w_{i+2}**: We put some frequent used words into single word groups to improve their discrimination ability.

4.3.6 Dependency on Previous Break

Although the simple CART approach works relatively well, it does not consider the dependency between breaks. Because each break is calculated separately, mistakes cannot be corrected using relationships between prosodic word breaks. Therefore, a model that can account for the dependency between the prosodic word breaks should be used.

As we are aware that the current break type (existence or not) is somewhat dependent on the previous break types. For example, if there is a prosodic break in previous position, the chance of the current position being a prosodic word break will

be less. In view of this fact, we can take the previous break type as one of the input features in our model.

4.3.7 Global Optimization

Using a window of four words and previous break, we can have a well prediction of prosodic word break. However, it is desired to consider longer dependency in the sentence. That is, we need to insert breaks into word sequence with global optimization. The following approach is proposed to have a better result of prosodic word prediction.

(1) Dependency Model

The approach for predicting prosodic word uses classification tree and a Markov assumption on the break sequence.

The probabilistic approach for the prediction of prosodic word breaks uses a stochastic model $P(a_1^n \mid Y_1^n)$ that represents the conditional dependence of the sequence of the breaks $a_1^n = \{a_1, a_2, ..., a_n\}$ on the sequence of feature vectors $Y_1^n = \{Y_1, Y_2, ..., Y_n\}$ for a sentence of *n* words (Ross, 1995). a_i is the type of the break (0 for non-break and 1 for break) after the syllable *i*, and Y_i is a vector of features that are relevant to the break. Using the chain rule:

$$P(a_1^n \mid Y_1^n) = p(a_1 \mid Y_1^n) \prod_{i=2}^{n} p(a_i \mid a_1^{i-1}, Y_1^n) \qquad (4.1)$$

Under the first-order Markov assumption, it is assumed that a_i is only dependent on a_{i-1} and Y_i which gives:

$$P(a_1^n \mid Y_1^n) = p(a_1 \mid Y_1) \prod_{i=2}^{n} p(a_i \mid a_{i-1}, Y_i) \qquad (4.2)$$

To calculate the $p(a_i \mid a_{i-1}, Y_i)$, CART approach is applied. a_{i-1} and Y_i are used as input features of the tree and $p(a_i \mid a_{i-1}, Y_i)$ is the output value of the tree. Normally

when using a decision tree, terminal nodes assign the most likely classification. Here, each node is associated with a discrete distribution, which represents the conditional probabilities for each break type. That is, we can obtain the values of $p(a_i = b \mid a_{i-1}, Y_i)$ from the tree, where b is a break type (break or no-break). The calculation of probability can be illustrated in Figure 4.1

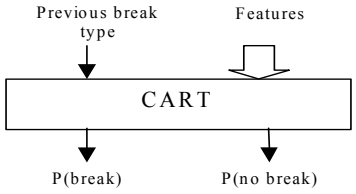


Figure 4.1 Prediction of probability using Classification tree

We want to determine the break between w_i and w_{i+1} in word sequence (w_{i-1}. w_i, w_{i+1}, w_{i+2}) around the break. The features (Y_i and a_{i-1}) used for the CART for probability calculation are:

- POSs of w_{i-1}. w_i, w_{i+1} and w_{i+2}
- Lengths of w_{i-1}. w_i, w_{i+1} and w_{i+2}
- Single word group of w_{i-1}. w_i, w_{i+1} and w_{i+2}
- Prosodic word break type of previous break position**.**

(2) Prediction Algorithm

The model tries to find the predicted break type sequence that maximizes the probability $P(a_1^n \mid Y_1^n)$. This can be achieved by using a Viterbi search algorithm. The prediction algorithm works as the following:

1. *Initial state P(0,1) = 1*
2. *Search*
   for i = 1 to N_{Word} *do*

a. *for k = 1 to* N_{Path}

for j = 0 to 1 do

$P(i, 2N_{Path} + j) = P(i-1,k)P(j)$

b. *sort P(i, j), (j=1 to* $2N_{Path}$*)*

c. *keep the first m items if* $2N_{Path} > m$

3. *Back trace to find the best break assignment sequence*

4. *Output*

where $P(i,j)$ is the probability of *jth* path in *ith* step, N_{Word} is the number of words in the sentence, N_{Path} is the number of paths in this step, $P(j)$ is the probability of the break type *j* in this step, *m* is the number of paths kept in current step (beam width of Viterbi Search).

(3) Model Training

The training process is to construct the decision tree and then calculate the probability of each class in a node.

The constructing process of the tree is a process that splits the training data into small sets. When a tree is constructed, each leaf node has a set of training vectors and a predicted value. We do not just take the classification value. To obtain a probability, we consider the distribution of the values in the node.

For example, given the value of a_{i-1} and Y_i, if we are going to determine the break type of a $_i$ (1 for break, 0 for no-break), the features (a_{i-1} and Y_i) trace the tree down to node T, and then the probabilities are calculated as:

$$p(a_i = 1 \mid a_{i-1}, Y_i) = m / n \qquad (4.3)$$

$$p(a_i = 0 \mid a_{i-1}, Y_i) = 1 - m / n \qquad (4.4)$$

where, *n* is the number of training samples falling into node *T*, in which *m* samples have break value 1.

To accurately calculate the probability on each leaf node, the size of the node should be large enough. The sizes of leaf nodes are controlled to have a minimum limit in the building process of the tree in this work.

4.3.8 Experiments

The corpus consists of 3609 Chinese sentences. There are totally 27293 word breaks, among which there are 17040 prosodic word breaks.

The text script of each sentence is automatically word segmented and tagged with POS types. Prosodic words are labeled manually according to the recorded speech. Please note that errors in word segmentation and POS tagging are kept in training data. The reason is that, if we use corrected data, the final model may be sensitive to wrong word segmentation and wrong POS types.

Experiments are performed by investigating: (1) the proper parameters for training of decision trees; (2) the performance of using different feature sets; (3) effect of number of POS categories (4) number of single word group. (5) performance difference between the different models (simple CART model and dependency model).

(1) Training Parameters

Before conducting all the experiments, we need to consider some relevant parameters.

During building the classification tree, one of the parameters needed is the stop size. This size determines the minimal size of nodes in a tree and controls the splitting process of building a tree. A too large value of stop size will lead to a tree that is not precise enough, while a too small value will result in the tree being over-trained to suit the training data. After some experiments, I find that, for my data, the stop size should be at least 7. To calculate the probability values in dependency model, we also

need the size of a node is not too small. In my experiments, I decide to use 20, which is suitable for all cases.

In tree construction process, we held some data for pruning (Refer to Section 2.3.3). We also investigate how much held out data should be used. This held-out data set is not used for testing of the result, but to build the tree. Therefore, it is part of training data. We investigated and found that there is no much difference when using 10% to 50% of the data as held-out. In the following experiments, I use 20% of the training data as held out pruning data.

Another problem is that how to divide all the data into training set and testing set. One can randomly select part of the data as testing data and the rest as training data. To get a more precise result, sometimes, 10-fold cross validation approach is used for training and testing. We compared our results using 20% randomly selected as testing data and 10-fold cross validation. The results are consistent. Therefore, in the rest of the work, we will use 80% of randomly selected data as training data and the rest 20% as testing data. We will concentrate on the features and schemes used for prediction in the following work. All the following testing results are results on testing set.

In all the following experiments, the trees are trained to maximize the accuracy of prediction of break and none-break. Accuracy is calculated as:

$$A = N_c / N_a \qquad (4.5)$$

where N_c is the number of correctly predicted break type (both break and non-break), N_a is the total number of all break type to be determined. All the following experiments will be evaluated using this value.

(2) Different Feature Sets

Now, we test how the model and features work when predicting prosodic word break. We use the information of two, three or four words to make the prediction. The result is shown in Table 4.4. In the table, Feature Pi and Li denote the POS type and length

of word Wi. We are going to determine the break types (existence or not) between W_i and W_{i+1}.

Features	Accuracy
P_{i-1}, L_{i-1}, P_i, L_i	73.68%
P_{i+1}, L_{i+1}, P_{i+2}, L_{i+2}	80.05%
P_i, L_i, P_{i+1}, L_{i+1}	84.01%
P_{i-1}, L_{i-1}, P_i, L_i, P_{i+1}, L_{i+1}	85.29%
P_i, L_i,P_{i+1}, L_{i+1}, P_{i+1}, L_{i+1}	85.37%
P_{i-1}, L_{i-1}, P_i, L_i, P_{i+1}, L_{i+1}, P_{i+2}, L_{i+2}	85.81%

Table 4.4 Accuracy of using different feature sets

In the table, we can see that using two words around the break to predict the break types achieves 84.01% of accuracy, which is better than using two words before or two words after the break. Using three words makes a better prediction, and using four words make the best prediction.

(3) Number of POS Types

The POS tagger of this work uses PKU tag set. Our corpus consists of 40 POS types (Categories). By merging some less frequent POS categories, we reduced it to 35 types. Because keeping too many POS categories may cause data sparseness problems, in this experiment, we will test if we can reduce the number of POS types without degrading the performance of the system.

There are two considerations of methods in reducing the number of POS (refer to 4.3.4). One is to merge the less frequent categories. The other is to merge the ones that contribute less to the break prediction. By training CART for prediction using POS of individual word as the only input feature (use P_i or P_{i+1}), we found that most of the less contributed ones are less frequent categories. That means the two options for reduction of POS categories come to the same way. We sort POS types in the descending order of frequencies in the data, and found POS types whose rank beyond 20 did not participate in tree construction. Therefore, we keep the most frequent 20 POS categories and merge all the rest into 1 category (the rest POS types).

We make the prediction using the features of POS (21 types) of four words and length of four words, and found the accuracy of prediction has not been affected (no

improvement and no degrade). We further reduce the POS categories by merging less frequent POS categories into the "Rest POS" class. We found that if the number of POS types reduces to less than 15, the accuracy starts to drop. Therefore, keeping 15 POS categories in the prosodic word prediction is sufficient.

We examine the most frequent types of POS category in the classification tree. It is interesting to find that three large POS categories do not participate in the classification. The three types are: v (verb), n(noun), p(prep). This can be explained that these three categories are actually a mix of words in different natures. They do not provide enough discriminating ability to the prediction.

(4) Number of Single Word Groups

We are aware that noun, verb, and adjective make very big POS categories. Therefore, some words cannot be well discriminated. In this experiment, we will form individual word group for frequently used word. The rest of the words will remain as one group. (Refer to Section 4.3.5) For example, if we decide to create 50 groups for the first 50 frequently used words, we will create 51 groups. 50 of them are for all the 50 frequently used words (Each word belongs to 1 group). All the other words belong to the 51st group. We take this word group value as a new feature in the test. By changing the number of the single word groups, we have the result shown in Table 4.5.

Number of word groups	Accuracy
0	85.2%
50	86.1%
100	86.2%
200	86.0%
500	85.8%
1000	85.6%

Table 4.5 Accuracy of different word group size

We find that when the number of word groups is 100, the accuracy is 1% higher than there is no word group defined. However, when the number of word groups increases, the accuracy begins to drop. The reason for this drop is that too many unnecessary categories will make the tree over-trained to suit the training data. Therefore, the number of word group should not be very large. Defining around 100

single word groups for the most frequent words would help to improve around 1% in prosodic word prediction.

(5) Dependency Model

In this experiment, we will compare the performance of dependency model with that of a simple CART model.

In the experiment, the features used are POS (20 POS types) and length of 4 words, and 50 single word groups. The results are shown in Table 4.6. In the table, simple CART approach achieves an accuracy of 86.10%. When previous break is added to input feature, accuracy improved to 88.10%. When using dependency model (previous break is considered in CART and constraints between breaks are considered in Markov chain), the accuracy improved to 91.65%.

Method	Features	Accuracy
CART without previous breaks	P_{i-1}, L_{i-1}, W_{i-1}, P_i, L_i, W_i, P_{i+1},$L_{i+1,}$ W_{i+1}, P_{i+2},$L_{i+2,}$ W_{i+2}	86.10%
CART with previous breaks	P_{i-1}, L_{i-1}, W_{i-1}, P_i, L_i, W_i, P_{i+1},$L_{i+1,}$ W_{i+1}, P_{i+2},$L_{i+2,}$ W_{i+2} B_{i-1}	88.10%
Dependency model	P_{i-1}, L_{i-1}, W_{i-1}, P_i, L_i, W_i, P_{i+1},$L_{i+1,}$ W_{i+1}, P_{i+2},$L_{i+2,}$ W_{i+2} B_{i-1}	91.65%

Table 4.6 Performance comparison for CART approach and Dependency model

Clearly, from the above results, we can see that the dependency model has better performance than the simple CART approach.

(6) Error Analysis

Our experiment result shows that the accuracy for prediction of prosodic word break on testing data is 91.65% (dependency model). 8.35% of the break types (existence or

not) are not correctly predicted as in the testing data. However, after analyzing the errors, the errors can be classified into two categories:

- Acceptable break type: Because there are many ways to break a sentence into prosodic word groups. Therefore, some predicted break types that do not agree with the testing data are actually alternative breakings. This accounts for about 3.4% of all the testing data.

- Unacceptable break type: Some others are wrong break types. The errors are mainly caused by wrong word segmentation and wrong POS types of words. Some are caused by ambiguity in sentence structure.

(7) Speed Comparison

We have proposed dependency model, which shows better performance than simple CART approach. We are also interested to know how fast the dependency model works. We conducted an experiment to compare the speed of dependency model and CART model. We predict prosodic word break of 4000 sentences and record the time used using the two different ways. In dependency model, we use a beam width of 30 (i.e. m = 30) in Viterbi search algorithm in Section 4.3.7. The test is done on a Pentium II-500 PC. The result is as shown in the Table 4.7. It shows that the speed of dependency model is around 36.0% of the CART approach.

Method	Time (seconds)
CART	9.0
Dependency model	25.0

Table 4.7 Speed comparison for CART approach
and Dependency model for prosodic word break prediction

4.4 Minor Phrase Break Detection

Phrase break is the break between phrases. There are many ways for break prediction in literature. The simplest methods usually only distinguish words as content words or function words; some use POS sequence. More approaches that are complex determine breaks based on a parsing tree. Due to the complexity of the parsing and the

low accuracy of parsing long sentences, this approach is not realistic in a real TTS system. In these approaches, some use POS sequence as input, each break is determined by a window of word sequence. However, the window size of word sequence used for prediction is limited by the data sparseness problems. In this work, we will try to overcome these problems using POS sequence but using CART to avoid data sparseness problem.

We examine the distribution of lengths of minor phrases. The distribution of the length is shown in Figure 4.2. From the figure, we can see that most of the minor phrases are within the range of 3 to 11 syllables. This means that minor phrase breaks are dependent on each other statistically. For example, a break is more likely to appear five to nine syllables away from its neighboring breaks.

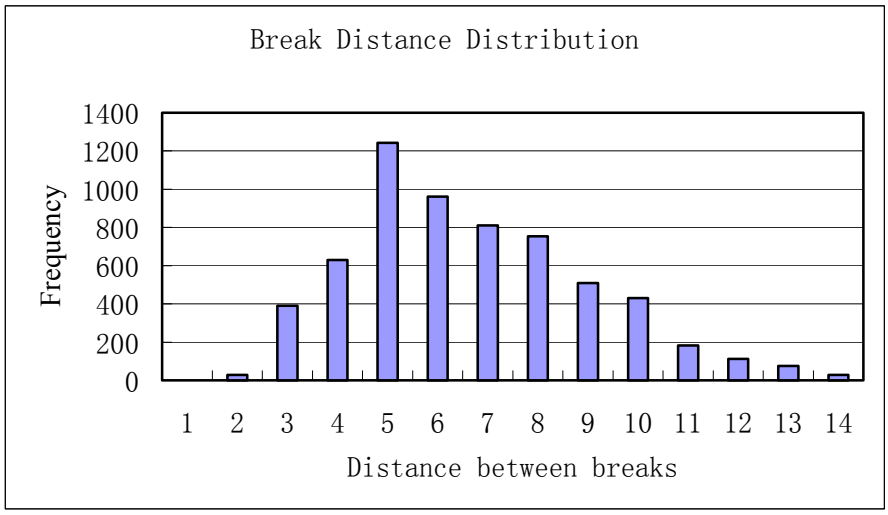


Figure 4.2 Distribution of number of syllables in phrase

We use break type in this section (Section 4.4). Break type means a binary value to indicate the existence of a minor phrase break. 1 means existence, while 0 means non-existence.

We calculated the mutual information between breaks and possible features. The result is shown in Table 4.8 and 4.9. Table 4.8 shows the mutual information (Refer to Section 2.4.1) between current break type and the previous break types. In the table, Break *Pi* means the previous ith break. Break P0 means the break to be determined itself. We can see in the table that, the highest mutual information value (0.00669) is

the value between Break P3 and the current break. That means a break type in a certain distance has some dependency relationship with the current break type.

Feature	Mutual Information
Break P7	0.00024
Break P6	0.00001
Break P5	0.00055
Break P4	0.00198
Break P3	0.00669
Break P2	0.00181
Break P1	0.00361
Break P0	0.98950

Table 4.8 Mutual information between break type and previous break type for minor phrase

Feature	Mutual Information
POS P8	0.00208
POS P7	0.00146
POS P6	0.00207
POS P5	0.00296
POS P4	0.01035
POS P3	0.02187
POS P2	0.04991
POS P1	0.22532
POS N1	0.23396
POS N2	0.15001
POS N3	0.01312
POS N4	0.00232
POS N5	0.00248
POS N6	0.00365
POS N7	0.00574
POS N8	0.00626

Table 4.9 Mutual information between break type and previous and next POS types for minor phrase

Table 4.9 shows the mutual information between break type and POS types of surrounding words. In the table, POS Pi means POS of the previous ith word, and POS Ni means POS of the next ith word. We can see that the highest values are for POS N1 and POS P1. That means that the POS types of words immediate next to the break have the highest influence on the break type to be determined. Words far away have less influence.

4.4.1 CART Approach

In a simple CART model, the break is mainly decided by the sequence of POS in the sentence. A window on the sequence can be used. If the size of the window for the current word i is from j words left from, to l words right from the word, then the features used for the prediction are: POSs of w_{i-j}, ..., w_{i-1}, w_i, w_{i+1},..., w_{i+l}. The determination of break type can be illustrated in Figure 4.3.

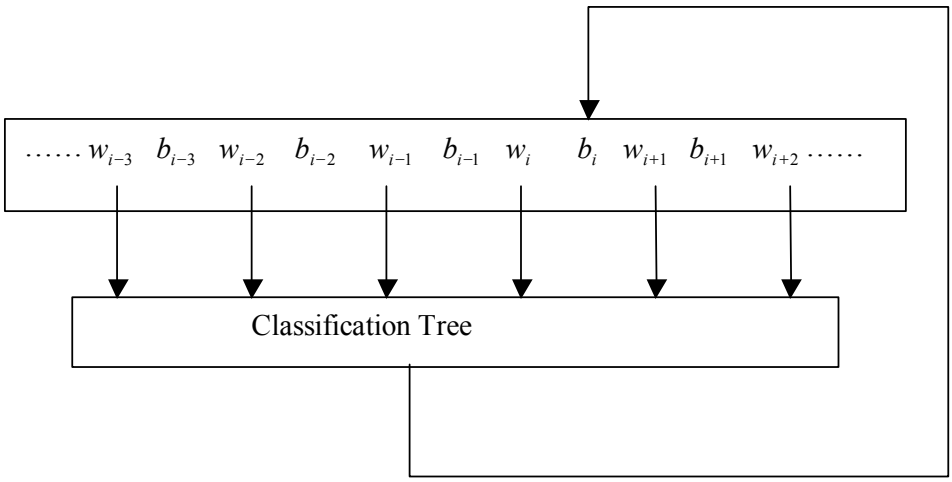


Figure 4.3 Calculation of probability using CART

4.4.2 Dependency Model

Inspired by the approach we used in prosodic word prediction, in this approach, we assume that the break between two words depends on the previous break sequence in this sentence. The model can be described as the following.

The probabilistic approach to prediction of minor breaks uses a similar stochastic model as for prosodic word detection. $P(a_1^n \mid Y_1^n)$ represents the conditional dependence of the sequence of breaks $a_1^n = \{a_1, a_2, ..., a_n\}$ on the sequence of feature vectors $Y_1^n = \{Y_1, Y_2, ..., Y_n\}$. a_i is the break type (break or no-break) of the syllable i, and Y_i is a vector of features that are relevant to the break. Using the chain rule:

$$P(a_1^n \mid Y_1^n) = p(a_1 \mid Y_1^n) \prod_{i=2}^{n} p(a_i \mid a_1^{i-1}, Y_1^n) \qquad (4.6)$$

Under the mth-order Markov assumption, the current break depends on m breaks before, we have:

$$P(a_1^n \mid Y_1^n) = p(a_1 \mid Y_1) \prod_{i=2}^{n} p(a_i \mid a_{i-1}^{i-m-1}, Y_i) \qquad (4.7)$$

To calculate the $p(a_i \mid a_{i-1}^{i-m-1}, Y_i)$, CART approach is applied. a_{i-1}^{i-m-1} and Y_i are used as input features of the tree and $p(a_i \mid a_{i-1}^{i-m-1}, Y_i)$ is the output value of the tree. a_{i-1}^{i-m-1} means the previous m break types, and Y_i means POS types of a window of word sequence around a break. The calculation of probability can be illustrated as shown in Figure 4.4.

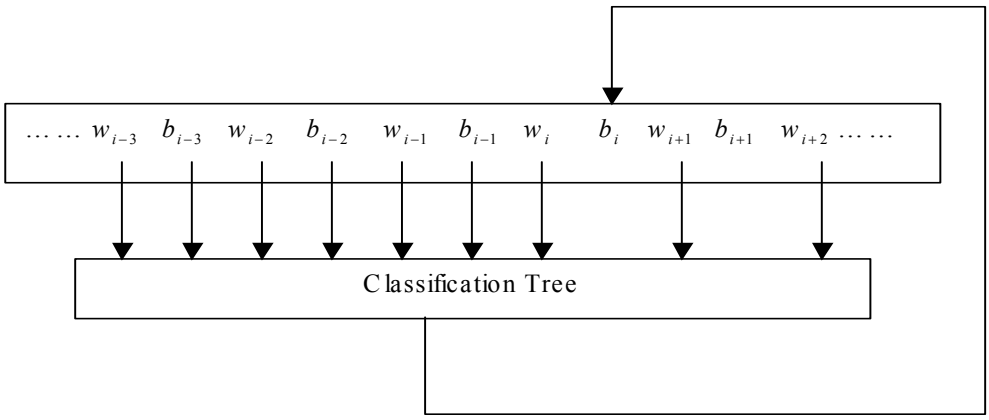


Figure 4.4 Calculation of probability using CART in dependency model

One more thing needs to be considered here. It is stated earlier in 4.2.1 that minor prosodic break set is a subset of prosodic word break. Therefore, the predicted minor phrase break cannot be in the middle of a prosodic word. To prevent this, the calculated probability should be adjusted. When a break position is not a prosodic word break, the probability value is assigned to zero. This avoids inserting a minor phrase break in the middle of a prosodic word.

The determination of breaks needs a dynamic programming process to find the best one. Viterbi search algorithm works similarly as that for prosodic word prediction.

4.4.3 Experiments

We use parameter precision (P) and recall (R) to evaluate the performances of the models. The parameters are defined as:

$$P = N_c / N_p \tag{4.8}$$

$$R = N_c / N_l \tag{4.9}$$

where N_c, N_l, and N_p are number of correctly predicted break, number of labeled break, and number of predicted breaks respectively.

Experiment is conducted to compare the dependency model and the simple CART model. We use the two approaches for testing:

CART approach with POS sequence (Method 1): In this approach, we only use POS sequence to predict breaks between words. A window consists of 2n surrounding words (n words before and n words after) around a word break. Features for predicting the break include the POS of n words before and n words after the break. For the cases that there are no enough words to fill the window, a NULL value is assigned as POS. In this experiment, n is given a value from 1 to 8.

Dependency model with POS sequence and previous break sequence (Method 2): In this approach, a window of 2n words is also selected together with n-1 breaks

before the words. Therefore the features for the prediction include POS of n words before the break, n words after the break and n-1 break types before the break. The n value also varies from 1 to 8.

The CART approach (Method 1) is previously used by other researchers. In this work, we take this approach as a reference to evaluate the performance of our new model, i.e. dependency model. Based on same corpus, the performances can be compared.

Similar to the experiments for prosodic words, we found that 20 is a suitable stop size for decision tree for our experiments. We used 10-fold cross validation approach in our experiment to better evaluation the models. The trees are trained to maximize the accuracy of prediction of break and non-break.

n	Recall	Precision
1	75.6%	70.9%
2	80.9%	74.5%
3	80.9%	74.2%
4	80.2%	74.4%
5	80.4%	74.5%
6	80.1%	74.2%
7	80.1%	74.4%
8	80.3%	74.9%

Table 4.10 Result of break prediction using CART and POS sequence

n	Recall	Precision
1	86.6%	75.1%
2	86.4%	80.2%
3	86.1%	80.9%
4	86.0%	80.0%
5	86.2%	80.5%
6	86.1%	80.4%
7	86.0%	80.5%
8	85.8%	80.7%

Table 4.11 Result of break prediction using dependency model

We calculated the precision and recall values for the prediction. The results of prediction are shown in Tables 4.10 and 4.11, and they are compared in Figure 4.5 and Figure 4.6.

(1) Performance of the Dependency Model

For dependency model, Figure 4.6 shows that precision increases from 75% to 80% when n changes from 1 to 2. There is no significant change when n > 2.

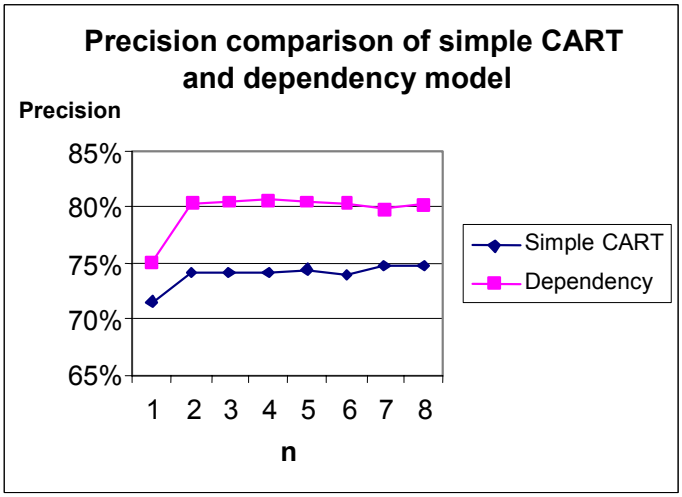


Figure 4.5. Comparison of precision values for phrase break prediction using the CART and dependency model

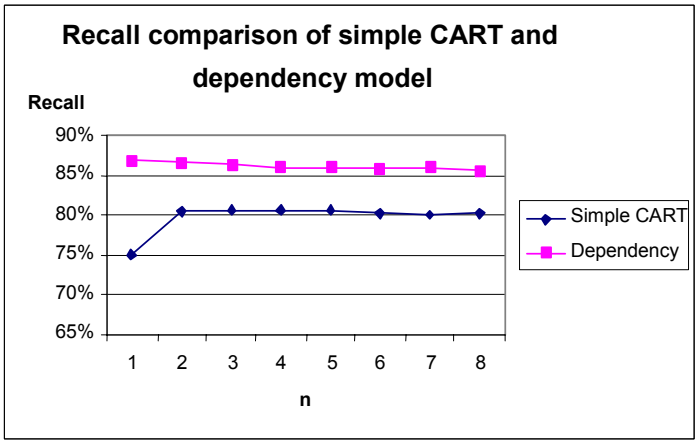


Figure 4.6. Comparison of recall values for phrase break prediction using the CART and dependency model

For recall values in Figure 4.7, there is no significant change when n changes from 1 to 8.

Therefore, the dependency model (when including 2 words before, 2 word after the break, and 2 breaks before the break) helps to improve precision. This does not have much influence on recall.

(2) Performance of the CART Model

In Figure 4.6, we can see that the precision value is around 71% when n=1, and increase to around 73% when n = 2. The precision values remain around 80% when n >2.

For the recall values in Figure 4.8, we can see it increases from 75% to 80% when n changes from 1 to 2. There is no increase when n > 2.

Therefore, it is necessary to include four words (two before and two after the break) into prediction when using CART approach for prediction.

(3) Comparison of the Two Models

In Figure 4.5, we can see then the precision values become stable when n>=2. The precision of dependency model is around 6% higher than simple CART model.

In Figure 4.6, we can see that there is 5% higher in recall values when n is greater than 2.

Therefore, when we include 4 words around the break for prediction, the dependency model has a better performance on both precision and recall than the simple CART approach.

(4) Error Analysis

Similar to prosodic word prediction, because there are many ways to break a sentence, the wrongly predicted breaks can be classified as acceptable break type and

unacceptable one. We find from our result that among all the breaks, around 3.5% of them are incorrectly predicted but are acceptable.

One of the drawbacks of the model is that some long minor phrases tend to be mistakenly separated into short minor phrases in our approach. This is due to, statistically, most of the minor phrases are short ones.

(5) Speed of Dependency model

We have conducted experiments to test the time of processing. The test is done on a PII-500 PC. We take 30 for beam width of Viterbi search. Experiment on 4000 sentences shows the speed of dependency model is around 40.1% (9.8/24.1) of the simple CART approach. The actual time used is as in Table 4.12.

Method	Time (seconds)
CART	9.8
Dependency model	24.1

Table 4.12 Speed comparison for CART approach and Dependency model for phrase break prediction

4.5 Discussion

From the experiments conducted, we have the following findings:

(1) For prosodic word break, we can achieve high accuracy when using four words around the break. This high accuracy shows that prosodic word break is dependent on the length and POS of the words around the break. Reducing the number of POS types to 15, there is no degrade for the performance. The performance can be improved by adding frequent single word categories.

(2) For prosodic word break, when applying dependency model, it shows better performance than using a simple CART approach alone. There is an increase of 5% in accuracy.

(3) For minor phrase break, when the dependency between minor phrase breaks is considered, there are increases in both precision and recall values in the same word

window size compared with simple CART approach. There are an increase of 6% in precision value and an increase of 5% in recall value.

(4) For minor phrase break, we find that both the CART approach and the dependency model achieves good performance when n = 2. Therefore, when making a prediction, it is necessary to consider two words before and two words after a word break, which is a 4-word window. For dependency model, we need to consider two word breaks before as well. There is no need to include more words.

(5) Directly comparing the performance of the work with other work is not easy because different experiments are based on different corpora, different features, etc. In addition, there is no public available corpus for testing different approaches for break prediction. However, compared with simple CART approach, which is used by many research projects, the dependency model has better performance in predicting prosodic word break and prosodic minor phrase break.

4.6 Summary

In this chapter, Chinese prosodic structure is first described. The problem of prosodic word and prosodic phrase has been investigated. Models for break prediction are proposed. Features for prediction are tested. Possible improvements are tried. Experiment shows that the proposed dependency model is better than simple CART approach.

For prosodic word prediction, we understand that length of words and part of speech are important features for Chinese prosodic word break prediction. There is a dependency between breaks, which will help to improve the accuracy of prosodic word break prediction.

For minor phrase prediction, the experiment shows that considering 4 words around a break can make a good prediction for both CART approach and dependency approach.

Chapter 5 Prosody Parameters

In this chapter, we investigate the problem of the prosodic parameters for unit selection based synthesis. First, we give an introduction of the prosody parameters and previous approaches in generation of prosody. Then, the definitions of the prosody parameters are given. Next, the parameters are evaluated and selected. Finally, the method for prediction of the parameters is given. Relevant experiments are described in the final part of the chapter.

5.1 Introduction

The naturalness of speech is determined by the richness of prosody contained in the speech. To generate high quality speech, proper prosody should first be generated from linguistic representation that is derived from an input text.

In a TTS system, the prosody is a set of parameters that describes rhythm, intonation, unit length, and loudness of speech. The values of parameters change with time. The main prosody parameters include the pitch contour of an utterance, duration of units, and energy of speech units. In the past decades, various approaches for predicting prosody parameters have been proposed for different languages.

For a given text, there are more than one spoken realizations by different speakers or for different intentions. The differences between the realizations might be very large. Lack of deep understanding of a sentence makes the determination of prosody of a sentence difficult. Usually, the resultant prosody is an average of or the most probable one among different possible realizations of the same linguistic expression. Therefore, the task of TTS usually generates the speech with commonly acceptable prosody for a text. Acceptable prosody means the prosody of the generated speech should be plausible but need not to be the most appropriate for a particular case (Monaghan 1989).

5.1.1 Pitch Contour

Pitch contour represents the change of fundamental frequency (F0) over time. It is generally accepted as the most important element of prosody. Various approaches have been applied in fundamental frequency generation. We can classify the existing approaches based on different aspects of prosody models. This part gives some characteristics of the prosody models.

(1) Direct Prediction or Two-step Prediction

The current TTS systems follow two general ways to generate prosody parameters. Some systems create prosody parameters directly from linguistic features. Other prosody models generate prosody in two steps: (1) a fisrt step to predict intermediate prosodic labels from text, and (2) a second step to convert the intermediate prosodic labels and other features into quantitative prosody parameters.

Usually abstract prosodic labels serve as intermediate prosodic labels. The abstract labels are designed according to prosodic theories of the language in research. Labeling systems, such as ToBI (Silverman et al 92) for English, are based on the perception of human. The labels capture the global intonation of a sentence and some important prosodic phenomena, such as pitch accents, boundary tones. The second step is a realization of the abstract labels.

(2) Parametric or Non-parametric Model

The parametric approaches try to describe the pitch contour with some parameters. In realization, pitch contour is generated by curve functions or interpolation. Typical parametric approaches include: Addictive model used by MITalk (Allen, 1987, O'Shaughnessy, 1979), Pierrehumbert's model (1981), Fujisaki model (1988). Fujisaki proposed a source/filter model to generate F0 contour. It defines two kinds of commands, phrase commands and accent commands. The former carries information about prosodic phrase and models as pulses. The latter represents a lexical accent and models as step functions. F0 contour is generated by smoothing the command signals with second order linear filters.

Non-parametric approaches, however, try to directly generate the final parameters from all the input features. The non-parametric approaches include hidden Markov model, neural networks, and concatenative methods. Most of corpus-based approaches are non-parametric.

(3) Tone-Sequence Model or Superpositional Model

It is a common knowledge that an F0 Contour is the result of many interacting factors, each having a different temporal scope (phone, syllable, word, phrase, sentence, or paragraph). A superpositional model attempts to model some or all of these factors separately, and combines the partial models to a final F0 contour (Buhmann et al, 2000). The final pitch contour is a combination of several contours. The famous Fujisaki model (Fujisaki, 1988) belongs to this category. A sequential model however directly generates F0 contour from left to right as a sequence of F0 values or movements. The tilt model (Taylor, 1998) and many other ToBI based models fall in this group.

(4) Rule–Based or Corpus-Based Approaches

Early systems use rule-based approaches (Klatt, 1987; Lee et al., 1989,1993; Chan and Chan, 1992; Anderson et al., 1984; Jilka et al., 1999). Currently, most prosody models have moved to corpus-based approaches. Typical corpus approaches include: CART approach (Lee S. H., 2000; Ross, K. N., 1995), Markov model (Ljolje and Fallside, 1986), linear regression (Black et al., 1996), neural networks (Traber, 1992), and others.

5.1.2 Duration

Duration means the time length of a speech unit. It is a way to describe the temporal structure of speech. Duration usually changes with many factors, such as phone identity, accent, phrase-final, etc. A duration model can predict duration for individual phoneme, or for a larger unit such as syllable.

The models to predict durations fall into two categories: rule-based and corpus-based. Currently, there is a trend to use data-driven approaches for duration modeling. Generally, there are two kinds of methods, which can be classified as parametric and non-parametric methods.

The most famous rule-based model is Klatt's model (1987) for English in MITalk. It used a multiplication formula. The parameters reflecting the contribution to durations were carefully tuned by researchers.

Van Santen (1994) proposed a sum-of-products model, which is a generalization of additive model and multiplicative model. The model is a sum of terms with each term itself is a product of one or more factors. The reported result of the model is that the correlation between observed and predicted duration was above 0.9 for both vowel and consonants.

Riley (1992) applied a CART approach to duration prediction.1500 utterances from one speaker were used to train the regression tree. The standard deviation of residual of prediction is 23ms.

5.1.3 Energy

Energy is considered less important than pitch contour and duration. Therefore, many systems do not treat energy seriously. However, inappropriate energy level of a unit may make speech sound uncomfortable. Therefore, full prosody control of speech needs to consider energy as well. Energy can be represented as a contour over time axis or a single value for a speech unit.

Corpus-based approach is generally adopted in generating energy contour. Neural networks (Lee et al, 1998), regression tree (Bagshaw 1998), and dynamic system (Ross and Ostendorf 1999) approaches were used to model energy contour.

The basic unit for energy prediction can be at syllable level (Lee et al. 1998), phone level (Bagshaw 1998) or even frame level (Ross and Ostendorf, 1999).

5.1.4 Previous Approaches for Chinese Prosody

For Chinese language, some models have been proposed to generate intonation contour, duration, and other parameters.

For pitch contour of Chinese, emphasis is put on two parts. One is F0 contour of lexicon tone. The other is global intonation of pitch contour. Rule-based systems model each tone with a contour and use a decline line to represent the global intonation. Lee et al. (1989) classified tone contour into some patterns, and rules were used to select different patterns. Bell labs system (Sproat, 1998) uses abstract labels to represent tones. Rules are defined to assign labels to syllables. The labels are further converted into pitch values.

Stem-ML approach (Shih et al., 2000) was proposed to model Chinese pitch contour. This is a parametric model, which can make quantitative F0 predictions, in terms of the lexical tones and the prosodic strength of each word. The model can accurately reproduce F0 in continuous speech with a 13 Hz RMS error.

For duration modeling, many attempts were made. Early systems determine the durations using handcrafted duration rules (Chiou et al. 1991; Choi et al. 1994). Parametric approaches were also used in Bell Labs Mandarin System (Shih and Sproat, 1996). Neural networks approaches were used by Hwang et al (1996), and Shih and Ao (1997).

There are a few corpus-based models for the generation of full prosody parameters. Neural network models (Chen et al., 1998) were applied to generate all prosody parameters (including pitch, duration, and energy).

5.2 Problems and Solutions

Although various ways have been used to generate prosody for Chinese, few of them are suitable for unit selection based approach. In this section, I describe the problems of prosody for unit selection, and provide my solutions.

5.2.1 Problems of Prosody for Unit Selection

Parametric representation of prosody: Prosody can be expressed in two ways. One is using symbolic representation. Another is parametric representation. Symbolic representations of prosody include tone, break, etc, which are abstract linguistic representations. Parametric representations include pitch contour, duration, and energy values. Symbolic representations are finally realized by parametric representations in real speech. Although prosody is considered as one of the most important factors of synthetic speech, prosody was not well handled in unit selection-based systems. Some of the previous systems used symbolic prosody in unit selection. This can only achieve limited success in naturalness because symbolic prosody representations are discrete values, which are only rough descriptions of prosody. Therefore, to better describe prosody in unit selection, there is a need to use parametric prosody representation in unit selection.

Parameters for unit selection: Previous Chinese prosody models only predict duration, energy value, and a curve to describe the pitch contour. The parameters are used in speech synthesis process by changing the speech signal. For example, in PSOLA synthesis, lengthening the speech (to change duration) is done by inserting more pitch periods; lifting the pitch value (to change pitch) is done by reducing the offset between the signals to be added up; or changing volume is done by amplifying the amplitude. However, in a unit selection-based approach, each unit has particular prosody parameters. The prosody parameters of the unit do not cover the total prosodic parameter space continuously. Therefore, during selection of units, there is a problem on how to measure the similarity between units. In consideration of this, we need parameters specially designed for unit selection-based approach.

Parameter definition: The main problems in prosody of current Chinese TTS systems include: rigid rhythm, inadequate pause, unclear tone, discontinuity in speech, sudden rising or lowering in pitch, too long or too short sound, etc. The specific reasons for these problems are:

- General prosody parameter: Inappropriate pitch, duration, and energy values will lead to sudden rising or lowering in pitch, too long or too short sound, etc.

- Implementation or representation of breaks: Inappropriate implementation or inappropriate parametric representation of breaks may result in rigid rhythm, inadequate pause.

- Implementation or representation of tones: Inappropriate implementation or inappropriate parametric representation of tones may result in unclear tone and unclear sound.

Although the prosody parameters are intended to describe all prosody aspects, simply selecting some basic prosody paraemters (duration, mean of pitch, energy) cannot effectively represent prosody. These parameters do not necessarily convey important perceptual information correctly. For example, it is unknown whether the tone and break information are correctly preserved in the parameters. We have to find an approach to solve the problem of realization of these perceptual effects.

Parameter selection: When many parameters are defined, there may be some redundancy. We want to select from them a small set of descriptive parameters that is sufficient but concise. This is a problem of parameter selection.

Feature analysis: There are many features (linguistic, phonetic, and break information derived from the input text) for prediction of prosody. To better understand the problem of prosody generation, we should investigate the relationships between the prosody parameters and the features for prediction.

Prediction model: We should decide a prediction approach for predicting the prosodic parameters.

5.2.2 Implementation of Perceptual Effects

We find that prosody implemented in final speech contains two kinds of information, which are:

- Implicit prosody: The intrinsic properties of speech that are required by segmental property of speech. These are basic prosody parameters, such as

duration, energy and pitch. For example, for a certain syllable, duration value should be in a proper range. If the duration is too small, it will sound bad.

- Explicit prosody: The properties that can be identified as perceptual prosody effect. The effect is usually represented by a combination of some prosodic parameters. For example, break information and tone information are perceptual effects. They may be described by a group of parameters.

The structure of the prosody prediction and implementation in this work is as shown in Figure 5.1. We understand the process of prediction of prosody and implementation of prosody in speech from three aspects. The three aspects can be considered as three transformation chains, which are entity chain, general prosody chain, and perceptual prosody chain. Note that, the three aspects are different understanding of the same process.

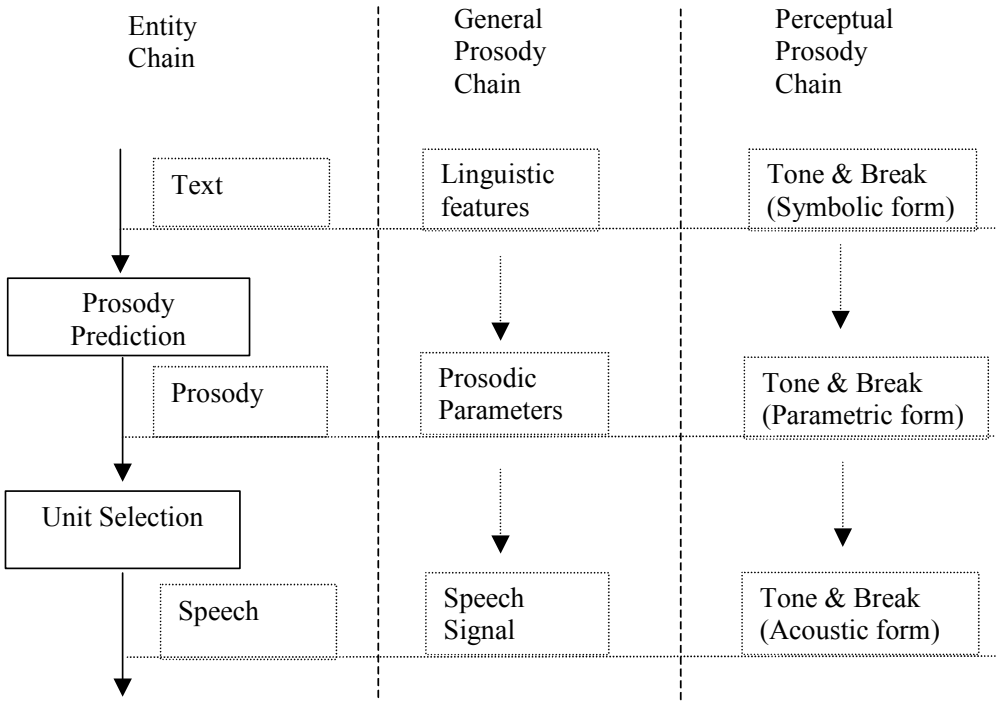


Figure 5.1 Prediction of prosody

In the entity chain, we see the prosody generation process transforms the text into prosody, and the speech synthesis process transforms the prosody (and other input) into speech. If we focus on general prosody properties, we can view the information transformation as the general prosody chain. In this chain, the prosody generation process transforms the linguistic features into the prosody parameters. Then the unit selection process generates a speech signal that contains the prosody parameters. From the view of perceptual prosody, we see that the tone and break information is contained in initial input text. After prosody generation, it is converted into a parametric representation. The parametric form of prosody representation is then converted into an acoustic representation after unit selection based synthesis process. We can see that the information of tone and break is transmitted in the whole process.

Therefore, from the entity view, the whole TTS process is to transform text into speech. From the view of general prosody, the text is transformed into proper speech signal with proper prosody properties. From the view of perceptual prosody, the identifiable perceptual elements (such as break and tone) are transferred though the prosody generation process and unit selection process to the final speech.

In this work, we want to determine the parameters that can correctly transmit the perceptual effects (e.g. tone and break through the chain). The process of determining prosody parameter set works as follows. First, an initial parameter candidate set is decided. Among the parameters, some of them should be sufficient to describe the desired perceptual effects. Then, the parameters are evaluated using two approaches. One is to examine the parameters from the statistical view to find their discriminating ability for the symbolic prosody representation. The other is to use recognition approach to verify the parameters. Properly designed parameter set can result in a sufficiently high accuracy. Next, a parameter clustering approach is used to select a set of units with minimal redundancy. Finally, the prosody parameters are integrated into cost function to guide the unit selection.

Note that, in our unit selection synthesis process, prosodic word break is implemented by selecting proper boundary syllables rather than inserting silences.

5.2.3 Solutions for the Problems

We give solutions to the problems raised in Section 5.2.1.

Parametric representation of prosody: In a unit selection-based synthesis approach, prosody parameters are used as discriminating criteria, which are used in a pattern matching process. Therefore, we decide to use some key parameters to describe prosody. We choose syllable as our basic unit for prosody analysis and generation. The calculation and prediction of the parameters will be on syllable level in this research.

Parameters for unit selection: In unit selection-based synthesis, the values of the parameters of a unit will be compared with the target values during unit selection process. When there is a mismatch, we should have a way to evaluate the degree of mismatch. In this work, we view prosody prediction as a classification problem. An input prosody feature vector will be mapped to a class. Each class has a predicted prosody parameter value, and a measure to account for the variation of the parameter. That means each predicted parameter would be represented by: (1) a value of the prosody parameter. (2) a variation measure of the predicted value. In this work, variation is measured using standard deviation of the samples in the same class.

Parameter definition: Tone and break are two of the most important prosody elements of Chinese speech. In this work, we will investigate the problem of describing effects of tone and break in speech. We will define parameters that are suitable for describing the tone and boundary effects. The defined parameters will be evaluated by statistical analysis and recognition.

Parameter selection: To remove redundancy in the defined parameters, we decide to use a clustering approach. The parameters will be clustered according to the correlation values between them. Representative parameters will be selected from each cluster.

Feature analysis: We will also examine all the factors that affect the prosody parameters. We are interested in which features are mostly affecting the result of

prosody parameters and which group of features can give a good prediction of the parameters. We will do the following:

- Prediction using single feature: We will find out the prediction ability of each feature we used.
- Prediction using stepwise training: We will find out which group of features have best prediction ability.

Prediction Approach: A prosody model is to map the linguistic input vector $L = (l_1,l_2 \ldots l_m)$ to prosody parameter vector $P = (p_1,p_2 \ldots p_n)$. Each p_i is a function of L.

$$p_i = F_i(L) \qquad (5.1)$$

where F_i is the function that derive parameter p_i from L.

To implement this function, we use CART approach. The inputs of the linguistic features are discrete values. The output p_i's are continuous values in this research. Due to the large number of features and training data items, the generated tree can be very large. The number of nodes may be hundreds or even thousands.

5.3 Prosody Parameters for Unit Selection

In this part, we will define a set of candidate prosody parameters to describe prosody for unit selection.

5.3.1 Duration and Energy

Duration means the time length of a unit. Duration of a unit is usually measured from start of the unit to the end of the unit. Start and end of a unit is labeled in the corpus. However, how to accurately determine the start and end of each syllable is a problem. We realized that duration actually relies on energy change. Start of a unit is identified when the energy value rises from zero up to a non-zero value, while end of a unit is identified when the energy value returns to zero. The problem is that sometimes energy is changing gradually. A unit may last too long before the energy goes to zero.

That makes the duration unstably long. To overcome this problem, in this work, we first normalized the duration of unit by removing low energy part.

During the calculation of duration, a normalization approach is used to obtain a consistent calculation of duration. The method of normalization is shown in Figure 5.2. The figure shows energy change of a syllable. The normalized s_n and e_n meet the following criteria.

$$F(s,e) = \int_s^e E(t)\,dt \tag{5.2}$$

$$F(s_l,s_n) = \alpha \cdot F(s_l,e_l) \tag{5.3}$$

$$F(e_n,e_l) = \beta \cdot F(s_l,e_l) \tag{5.4}$$

where, *E(t)* is the RMS energy of the signal at time t and $F(s,e)$ is the accumulative energy from time *s* to time *e* as illustrated in the figure. s_l and e_l are labeled start and end. α and β are small values, e.g. 0.001. By using this processing, silence parts or the parts with very small sound are excluded from the duration of syllables. As duration is only served as criteria for unit selection, it does not hurt even if part of the unvoiced initial of a syllable is excluded from the duration.

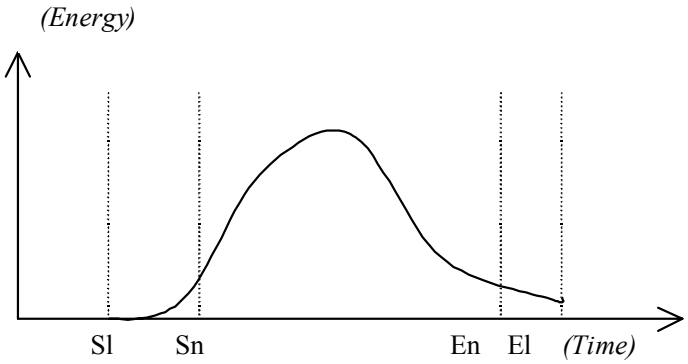


Figure 5.2 Syllable duration normalization

By using this normalizing approach, in the corpus, the mean of standard deviation of duration reduces from 65.9 to 64.6 ms (with mean from 243.3 to 240.7 ms). We

examine the change of durations, and found that 12% of the units have more than 0.01 second change in duration. For all the changes, we found that most of them are start or end syllable of an utterance. The start and end silences have been removed from duration. We use the normalized duration as the syllable duration parameter.

Energy is a parameter to measure the loudness of sound. There are a number of representations of Energy (and they might be in different scales, e.g. dB which is in logarithm scale). For example:

- Total value:

$$E = [\sum_{i=1}^{n} x^2(i)] \tag{5.5}$$

where n is the sample number in a unit, *x(i)* is the signal value of the ith sample.

- Maximal value

$$E = \underset{i=1}{\overset{n}{Max}} (E_{RMS}(i)) \tag{5.6}$$

$$E_{RMS}(i) = \sqrt{[\sum_{j=1}^{m} x^2(i - m/2 + j)]/m} \tag{5.7}$$

where *n* is the number of syllables in a unit, *m* is a frame length for calculating RMS energy, *x(i)* is the signal value of the ith sample.

The two kinds of representation (sum value or maximum value) do not consider influence of duration for energy. The total energy of unit reflects energy over all the duration of whole unit. For same type of unit, a unit with long duration usually has higher total energy than that with a shorter duration. Maximal value of RMS Energy reflects the peak value of energy in the unit. It only reflects part of the energy information of a unit. A better way is to use an average value of energy within the duration of a unit.

As we know that energy of unvoiced part of a unit is low, including unvoiced part into energy measure may introduce unstableness in energy value. In this work, we use RMS energy of syllable only on voiced part of the unit. The RMS energy is defined as:

$$E = \sqrt{[\sum_{i=1}^{n} x^2(i)]/n} \qquad (5.8)$$

where *x(i)* is the amplitude of the ith sample of the signal, *n* is the number of samples in the voiced part of the syllable.

Duration value and energy are important element of prosody. However, the values (even with pitch parameters also considered) cannot fully reflect some important difference in prosody. For example, the parameters cannot distinguish boundary syllables. A start syllable and an end syllable of prosodic word have markedly different perception effects. Incorrect use of boundary units will result in wrong break position effect in speech utterance. So we have to investigate the acoustic correlates of boundary units.

Energy contour is one of the options for this consideration. However, description of energy contour depends on the start and end markings of a unit, while the start and the end of the unit depend on energy contour (i.e. at the edges of unit, what an energy value can be considered as silence). To solve this paradox, in this work, we use a representation by considering energy and duration simultaneously.

The new defined parameters are based on the Figure 5.2. Similar to formula 5.2, we define parameters using the following formulas.

$$F(s_{n,}s_{\gamma}) = \gamma \cdot F(s_n, e_n) \qquad (0 < \gamma < 1) \qquad (5.9)$$

$$p(\gamma) = (s_{\gamma} - s_n)/(e_n - s_n) \qquad (5.10)$$

where γ is a given value for defining parameter, s_{γ} is the corresponding time point within the duration, $p(\gamma)$ is the defined parameter, others have the same meaning as those in formula 5.2.

$p(\gamma)$ defines a percentage point of energy distribution in the duration. It is another description of energy contour. Take an example to explain the meaning of γ and $p(\gamma)$. If we define $\gamma = 0.3$ and calculated that $p(\gamma) = 0.4$, it means that divided by s_γ, the left part of the syllable accounts for 30% of the energy and 40% of duration.

We set values of γ, and calculate values of $p(\gamma)$ as our prosody parameters. In this work, we define percentage points of duration that divide energy at 1/6, 2/6, 3/6, 4/6, and 5/6 of whole energy. That is, γ takes 1/6, 2/6, 3/6, 4/6, and 5/6 in formula (5.10).

Besides the parameters we defined above, we should define two other parameters, which are parameters that describe energy level at boundaries. It is usual that the energy value at syllable boundary is not a value close to zero. Rather, in many cases, because a unit is tightly connected with previous or next units, there are continuous energy contour between two units. Therefore, we represent the boundary energy (start and end position of a unit) with RMS values within a 50 ms frame.

5.3.2 Pitch Contour

Pitch contour is generally considered as the most important one among prosody descriptions. In this research, pitch contour is decomposed into two parts. The pitch contour is considered as the sum of global intonation contour and syllable pitch contour.

- **Global intonation contour:** Global intonation contour means the global change of pitch values over the syllables in a sentence. It controls the whole intonation of an utterance. The global contour is determined by the grammatical function and pragmatic function of each word and phrase in the sentence.

- **Syllable tone contour:** Syllable F0 contour means the local change of pitch values in a syllable. It controls the tone identity of a syllable. Syllable contour

is usually determined by tone of the syllable, and affected by tones of surrounding syllables, stress degree, etc.

Suppose the F0 contour for the voice part of a syllable is $f(t)$ and s and e are start time and end time of the voiced part of the syllable. Then we define the following:

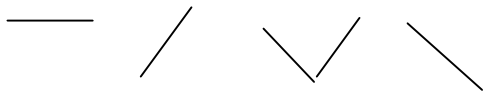
Pitch Mean of Syllable: mean pitch value of a syllable

$$p = (\int_{s}^{e} f(t)dt) / (e - s) \tag{5.11}$$

Tone Contour of Syllable: Tone contour is defined as the pitch contour of a syllable minus the pitch mean of the syllable.

$$c(t) = f(t) - p \tag{5.12}$$

where p is pitch mean of the syllable.



Tone 1 Tone 2 Tone 3 Tone 4

Figure 5.3 Illustration of pitch curves of tone

Tone Contour Vector of a Syllable: Tone contour is expressed using a vector. We obtain m samples in the pitch contour evenly to form an $m+1$ dimensional vector. This gives a uniform representation of all syllable pitch contour. Tone contour vector of the syllable is defined as:

$$C = \{c_0, c_1, c_2, ..., c_m\} \tag{5.13}$$

$$c_j = f((j-1)\Delta t), j = 0..m \tag{5.14}$$

$$\Delta t = T / m \tag{5.15}$$

where *T* is the duration of the voiced part of the syllable.

The global contour can be expressed by pitch mean values of syllables of sentence. To express contour the local of each syllable, we use the tone contour vector. In this work, m takes value 8. To more efficiently describe the contour of tone, we need to define more parameters.

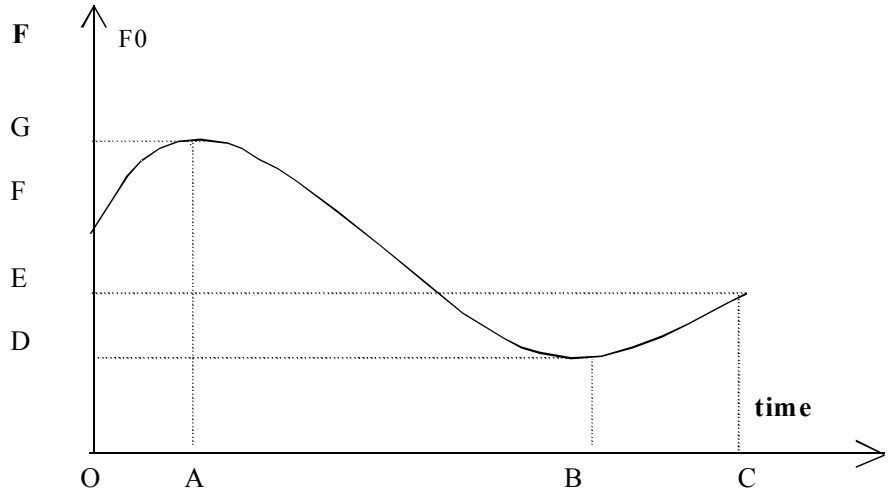


Figure 5.4 Illustration of prosody parameters

Before defining more parameters to express local tone contour, we have a look at the stylized pitch curves of four tones in Figure 5.3. We can easily see that each tone has clear difference in start and ends. Therefore, we use parameters to characterize these values. Former research also shows that pitch range is an important factor for Chinese prosody (see 2.2). Therefore, the parameters to characterize local contour of a syllable are defined as following (Refer to Figure 5.4):

PitchRange: The difference between the maximal value and the minimal value of pitch contour. (DG in the figure).

PitchStart: The pitch value of the start point of the voiced part. (OF in the figure).

PitchEnd: The pitch value of the end point of the voiced part. (OE in the figure).

5.3.3 Candidate Prosody Parameters

A summary of all defined parameters for each syllable is as the following:

1. **Duration:** The time length of the syllable.

2. **EnergyRMS, EnergyMax, EnergySum:** Average, Maximum, and Sum of energy of the voice part of the syllable. EnergyRMS is the RMS energy within the whole voice part of the syllable.

3. **PitchMean:** Mean value of pitch of the voiced part of syllable.

4. **PitchRange:** The difference between maximal value and minimal values of pitch contour in a syllable.

5. **PCon0, PCon1, PCon2, PCon3, PCon4, PCon5, PCon6, PCon7, PCon8:** The values are defined in formula (5.13) when m takes 8. The reason of using 8 is that, after examining pitch contours of syllables, I find sampling 8 points is enough to describe the main shape of the pitch contours. In all the values, for the convenience of later use, we also represent PCon0, PCon4, PCon8 as PitchStart, PitchMiddle, PitchEnd respectively, which are just values of the start point, middle point and end point of the voiced part.

6. **EnergyStart, EnergyEnd**: RMS energy values with a frame of 50 ms at start and end points of each syllable.

7. **EnPer1, EnPer2, EnPer3, EnPer4, EnPer5:** The values describe 5 percentage points within the duration. The 5 points divide the whole energy of the syllable equally into 6 segments. That is, γ takes 1/6, 2/6, 3/6, 4/6, and 5/6 in formula (5.10). Here, we divide duration into 6 segments because we find it is enough to describe the trend of an energy contour. EnPer3 is also represented as EnergyHalfPoint, for the convenience of later use.

Among the parameters we defined, each has its main concerns:

- Duration, Energy, and PitchMean are general parameters that determine the global prosody of utterances (although they also have effects on local prosody).
- EnergyStart, EnergyEnd, EnPer1, EnPer2, EnPer3, EnPer4, and EnPer5 together with duration are mainly used to describe boundary effects (i.e. break).
- PCon0, PCon1, PCon2, PCon3, PCon4, PCon5, PCon6, PCon7, PCon8 together with PitchMean and PitchRange are mainly used to describe tones.

5.4 Parameter Determination

In all the candidate parameters, some are intended to express the perceptual effects, such as break and tone. We will evaluate the parameters to see whether they effectively express the effects. Then redundancy will be removed and a concise set will be selected.

5.4.1 Parameter Evaluation

We have defined the parameters to describe prosody. However, one problem is: are these parameters sufficient to describe important aspects of Chinese prosody? Two most important prosody properties of Chinese speech we are to realize in speech synthesis are tone and break (prosodic break). Therefore, we will examine whether the defined parameters are fit for describing them. To simplify the work, we only consider prosodic word break. Therefore, break means prosodic word break in this context. We will investigate this by:

- Examining the distribution of the parameters for different tone types and boundary types. We will use boxplots to see the parameters have different distributions for different types of tone and boundary type. By using this way, we make sure that the parameters we will use are relevant parameters to the intended prosodic effects.

- Examining the ability of the parameters for describing tones and breaks from the view of tone and boundary recognition. If a computer can recognize the tones correctly, it is possible that human can easily perceive the tone based on the acoustic properties of the speech. We will use CART approach for the recognition purpose in the work. By using this way, we make sure that the parameters we will use are sufficient to describe the prosodic effects.

Details of the parameter evaluation will be described in experiment part in Section 5.6.1.

5.4.2 Parameter Selection

We have listed all candidate prosody parameters in 5.3.2 and have confirmed that the defined parameters can describe tone and break in 5.4.1(details in 5.6.1). However, with so many parameters, it is not efficient to predict all of them because many of them are highly correlated. Therefore, we should choose some representative parameters from all the candidates.

In this work, we use clustering approach to reduce the number of useful parameters. The distance between parameters is calculated based on correlation value between two parameters.

We use absolute correlation distance in the work. For the absolute correlation distance method, distance is defined as:

$$d_{i,j} = 1 - |r_{i,j}| \tag{5.16}$$

where $r_{i,j}$ is the Pearson product moment correlation (Refer to Section 2.4.2) between variables i and j.

In this work, the distance between two clusters is the average distance between a variable in one cluster and a variable in the other cluster. The distance is defined as:

$$D_{k,l} = \left(\sum_{i=1}^{N_k}\sum_{j=1}^{N_l} d_{i,j}\right)/(N_k N_l) \tag{5.17}$$

where, N_k and N_l are the number of variables in clusters k and l.

The clustering process can be shown by a dendrogram. Then we cut the dendrogram at a similarity level and the clusters are determined. Final parameters are determined by choosing one parameter from each cluster.

Details of parameter selection will be described in experiments at Section 5.6.1.

5.5 Prediction of Prosody

5.5.1 Features for Prediction

Prosody is determined by many factors. The following features are defined as determining factors of prosody parameters in this research. All these features are input values in prediction.

(1) Syllable Information

Syllable information includes the syllable itself and its context syllable. Each syllable is a combination of initial, final and tone. There are following features:

- Initial of the current syllable (CurrInitial).
- Final of the current syllable (CurrFinal).
- Tone of the current syllable (CurrTone).
- Initial of previous syllable (PrevInitial).
- Final of previous syllable (PrevFinal).
- Tone of the previous syllable (PrevTone).
- Initial of the next syllable (NextInitial).
- Final of the next syllable (NextFinal).
- Tone of the next syllable (NextTone).

(2) Word Information

Three words are considered for a syllable as possible determining factors. They are the word containing the syllable, previous word and next word. Each word has a length and a POS category. The features are:

- Length of the current word (WordLen).
- POS type of the current word (WordPOS).
- Length of the previous word (PrevWordLen).
- POS type of the previous word (PrevWordPOS).
- Length of the next word (NextWordLen).
- POS type of the next word (NextWordPOS).
- Location of the syllable in a word (LocInWord).
- Start syllable of a word (WordStart): 1 for Yes, 0 for No.
- End syllable of a word (WordEnd): 1 for Yes, 0 for No.

(3) Prosodic Word Information

Word is unit defined from syntax view. In speech, prosodic word is a more stable unit than word. The prosody of the syllable being the first syllable of a prosodic word is different from those of syllables in middle or final position of a prosodic word. Therefore, in this research, prosodic word is applied as a feature. The features are:

- Length of the prosodic word (PWLen).
- Tag indicating whether it is the first syllable of the prosodic word (PWStart). The value is 1 for yes and 0 for no.
- Tag indicating whether it is the final syllable of the prosodic word (PWEnd). The value is 1 for yes and 0 for no.
- Location of the syllable in prosodic word (LocInPW).

(4) Phrase Type and Breaks

Phrase is important in that (1) the boundary syllable is usually different from other syllables in prosody. (2) There is a decline trend for pitch in an utterance. In this work, we use features to indicate whether the syllable is a boundary syllable. We define the following features about phrase.

- Major phrase type (IPType): Major phrase is equivalent to intonation phrase in this work. Major phrase type is approximated by using type of utterance. The defined types are: (1) Incomplete utterance. (2) Statement utterance. (3) Questioning utterance.
- Location of the syllable in major phrase (LocInIP).
- Break type before the syllable (BreakBefore). The types include: No-break, word break, prosodic word break, minor phrase break, major phrase break.
- Break type after the syllable (BreakAfter). The types are the same as BreakBefore.

5.5.2 Prediction Ability of Features

(1) Single Feature in Prediction

As all the input information will be used as discriminating factors in our model to give accurate prediction using CART approach, we first examine the discriminating ability of each feature. This evaluation is done by using only one factor as classification feature and judge the accuracy of the classification made by this factor. For example, the tone of a syllable is one of the factors that affect the duration of the syllable. To find out to what extent the tone can be used as classification criteria, we classify the syllable into five classes by tone. In each class, we take the average value of the durations. Then we have five values, which will be used as the predicted value of the syllable. Comparing the predicted value with the actual value we obtain from corpus, we have a correlation between the two sets of values. This correlation will serve as an index of the distinguishing ability of the feature tone.

In this research, we examine the relationship between the features and the main parameters. But it should be noted that:

1. The result is a statistical result that reflects the corpus. Conclusions based on this corpus are true for the corpus within the same domain.

2. Some of the features are different representations of the same fact. Therefore, features may be dependent each other in this study.

3. The relationship between the input features and the output parameters may be cause-consequence or just statistical co-occurrence.

In statistics, to draw conclusion from one sample data sometimes is not reliable. In practice, these two techniques can make conclusion more reliable. (1) When sample size is small, for example, less than 50, a widely used approach is bootstrapping, in which data are re-sampled and statistics are based on many rounds of sampling. (2) When sample size is large, a typical approach is to randomly divide the sample into two disjoint sets. Statistical results from the two sets will be compared to make sure the result is consistent. Since we have a large corpus, we use the later as our preferred approach.

We conduct this experiment in Section 5.6.2.

(2) Combined Features in Prediction

In this part, we examine the prediction ability of combined features. This is done by using stepwise training of regression tree. In stepwise training of decision tree, each single input feature is considered in each step and the feature that can achieve the largest reduction in impurity is selected as a new feature in each step. By this way, a group of features that can contribute most to the training process are adopted first. The input features will be selected one by one by the order of importance in constructing the tree. Therefore, this part is to find a sequence of most important features that can give best prediction of a single prosodic parameter.

We conduct this experiment in Section 5.6.3.

5.5.3 Prediction Model

We are designing prosody for unit selection-based approach. One of the important factors in measuring unit mismatch is the degree of variations of a unit. There are two reasons why we need to consider variations of parameters. (1) Different parameters have different measuring scales. Without normalization, they cannot be compared together. (2) We are aware that even with same type of parameters, in different situations or for different unit identities, they have their own variation trends. For example, for energy of syllables in our corpus, syllables with final A have larger variations (standard deviation is 822) than those with final UN (standard deviation is 609). Therefore, we view prosody prediction as a classification problem. Feature vectors will be classified into classes. In each class, we calculate standard deviations, which will be the measure to account for the variations of the predicted parameters. CART approach can be used for classification and prediction. It is a natural choice to use it.

Each parameter we defined for this work is a continuous value. For each parameter, a regression tree will be built. Given all the feature values of a syllable, the regression tree will give a predicted value together with a standard deviation of the predicted value. The predicted value is the parameter value we expect, while the standard deviation describes how accurate the value might be.

We conduct experiments on parameter prediction in Section 5.6.4.

5.6 Experiments

5.6.1 Parameter Determination

In this part, we first conduct experiments to evaluate the parameters for describing tones and breaks. Then we cluster parameters to select a set of useful parameters. Finally, we will look at the properties of the selected parameters.

(1) Parameters Describing Tone

It is a common knowledge that the acoustic correlate of tone is pitch contour of a syllable in speech. PitchMean, PitchRange, Pcon0 (PitchStart), PCon1,Pcon2, Pcon3, Pcon4 (PitchMiddle), Pcon5,Pcon6, Pcon7, and Pcon8 (PitchEnd) describe pitch values. Therefore, they are parameters to describe tone.

First, we evaluate the discriminating ability of the parameters for tone types. We draw boxplots for this purpose. Among all the parameters, we draw boxplot of four parameters. Figure 5.5, 5.6, 5.7, and 5.8 show the boxplots for PitchMean, PitchRange, PitchStart, and PitchEnd respectively. In all the figures, tone 5 means neutral tone.

In Figure 5.5 for PitchMean, we see that tone 1 and tone 4 have a clear distinction from other tones in median, Q1 and W3. In Figure 5.6 for PitchRange, we can see that tone 1 and tone 4 have a clear distinction from each other. In Figure 5.7 for PitchStart, we can see that tone 1 and tone 5 have distinction between tone 3, tone 4, and tone 5. In Figure 5.8 for PitchEnd, we can see that tone 2 and tone 4 have clear distinction from each other. In brief, each of the four parameters provides some distinction between some tone types. Examination of the rest of parameters gives similar conclusion. Therefore, the parameters are useful in describing tone types for Chinese.

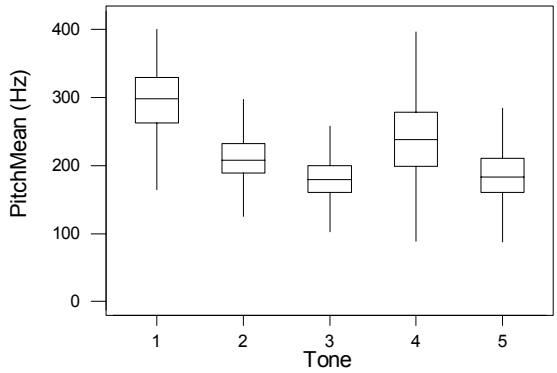


Figure 5.5 Boxplots for PitchMean by tone type

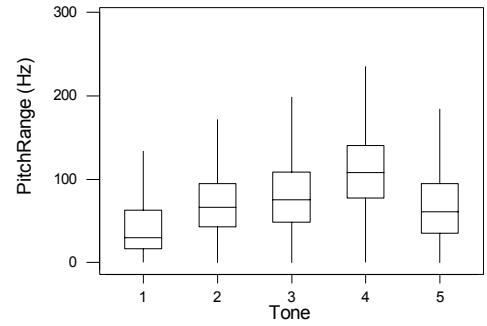


Figure 5.6 Boxplot for PitchRange by tone type

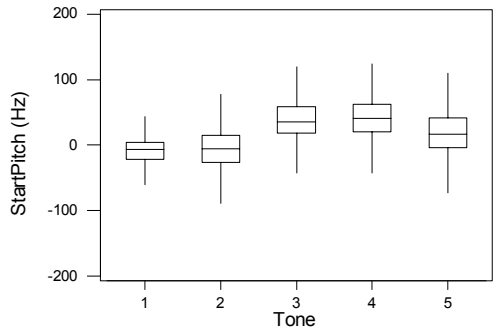


Figure 5.7 Boxplots for PitchStart by tone type

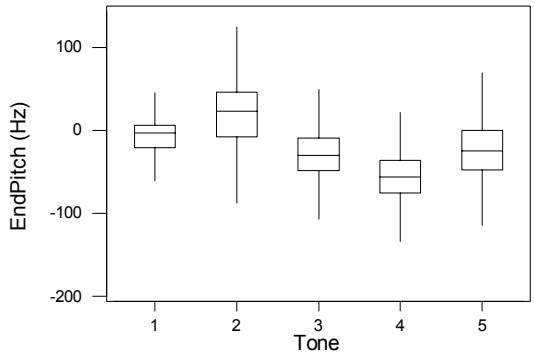


Figure 5.8 Boxplots for PitchEnd by tone type

To further evaluate the parameters, we use the parameters together with other parameters to predict the tone category. In other words, we are trying to recognize tone type based on the above parameters and other possible input. Again, we use CART approach for the recognition. The inputs of the classification tree are continuous values, while the output of the tree is tone type. The features for the recognition of tone in this investigation is as the following:

- PitchMean : Mean value of pitch.
- PitchRange: Range of pitch value for the tone contour.
- Nine sample values from pitch contour: PithchConP0 (PitchStart), PitchConP1, PitchConP2, PitchConP3, PitchConP4, PitchConP5, PitchConP6, PitchConP7, PitchConP8 (PitchEnd).
- EnergyRMS: RMS energy.
- Duration: Duration of the syllable.

Training Data							
Actual Class	Total Cases	Percent Correct	1 N=6162	2 N=7486	3 N=6209	4 N=8694	5 N=3420
1	6,027	93.8	93.8	2.5	0.8	1.3	1.5
2	8,156	84.9	2.7	84.9	5.3	0.8	6.4
3	5,656	74.5	1.5	3.9	74.5	8.5	11.6
4	10,190	77.7	1.6	1.2	13.4	77.7	6.2
5	1,942	78.4	2.1	3.5	7.9	8.1	78.4
Testing Data							
Actual Class	Total Cases	Percent Correct	1 N=6125	2 N=7510	3 N=6498	4 N=8519	5 N=3319
1	6,027	91.5	91.5	3.3	1.5	1.8	2.0
2	8,156	82.4	3.2	82.4	6.4	0.9	7.1
3	5,656	66.7	1.9	5.4	66.7	10.6	15.3
4	10,190	73.3	1.7	1.2	16.0	73.3	7.9
5	1,942	48.8	3.6	8.7	25.3	13.6	48.8

Table 5.1 Accuracy for tone recognition

Using CART approach with 10-fold cross validation, the result is as shown in Table 5.1. The table shows the accuracy of tone types. We can see in the table that:

- The lowest accuracy for testing test is for tone 5 (neutral tone). The accuracy of tone 5 for training data is 74.5%. However, for testing data, the accuracy is

only 48.8%. The reason for this low accuracy is that tone 5 is not a formal tone and there is not typical pitch contour shape for tone 5.

- If we ignore tone 5, the general accuracies for all the other tones are from 74.5% to 93.8% for training data, and from 66.7% to 91.5% for testing data. Therefore, except for tone 5, the general accuracy is quite good.

- The accuracy for training and testing data for tone 3 is low (74.5% and 66.7%). This shows that tone 3 is difficult to be correctly recognized. This accuracy is consistent with our observation that many of the tone 3 syllables are not clearly recognized by human ears.

- From accuracy of testing data, 16.0% of tone 3 syllables are recognized as tone 4 syllable, and 10.6% of tone 4 syllables are recognized as tone 3. This shows that tone 3 and tone 4 are sometimes difficult to be distinguished from each other. This is also observed during manual verification process of the corpus.

We calculate the total accuracy for all syllables, and find the accuracy of training data is 82.0% and that of testing data is 76.4%. If we ignore tone 5, the accuracy of training data is 82.3% and 78.2% respectively.

	Pitch Mean	Pitch Range	Pitch Con0	Pitch Con1	Pitch Con2	Pitch Con3	Pitch Con4	Pitch Con5	Pitch Con6	Pitch Con7
PitchRange	-0.105									
PitchCon0	-0.217	0.459								
PitchCon1	-0.147	0.541	0.920							
PitchCon2	-0.054	0.561	0.686	0.895						
PitchCon3	0.061	0.438	0.296	0.551	0.846					
PitchCon4	0.191	0.033	-0.378	-0.204	0.157	0.636				
PitchCon5	0.179	-0.419	-0.822	-0.864	-0.721	-0.312	0.513			
PitchCon6	0.098	-0.522	-0.790	-0.930	-0.955	-0.749	-0.028	0.824		
PitchCon7	0.048	-0.546	-0.696	-0.860	-0.951	-0.865	-0.314	0.565	0.910	
PitchCon8	0.013	-0.523	-0.572	-0.723	-0.828	-0.806	-0.416	0.341	0.710	0.912

Table 5.2 Correlation values between parameters for tone

To understand the accuracy, we conduct a listening test for 200 syllables by 3 persons. Each person is asked to listen to the 200 syllables and to count the number of

tones that can be clearly identified. The result shows that the average percentage of syllables with clear tone is 85.4%. This shows that the accuracy by tone recognition is close to the result of human perception. Therefore, the defined parameters can well describe tone.

We calculate the correlation values of the defined parameters for tones. The values are as shown in Table 5.2. From the table, we see that PitchMean has relative small correlation values with other parameters. PitchRange has moderate correlation values with others except for PitchCon4. The correlation values between PitchCon0 to PitchCon8 are diverse. Some are high and some are low. Generally, we can conclude that there are many redundant parameters in all the defined parameters for describing tone. We will remove the redundancy later in this chapter.

(2) Parameters Describing Break

Among the prosodic break types, the prosodic word break is the smallest prosodic break type and the biggest break set. In this part, we examine the parameters that are meant to account for the breaks. We know that at boundary of prosodic units, there are usually lengthen effects. This may lead to a longer duration for a syllable than at non-boundary positions. We define parameters Duration, EnergyStart, EnergyEnd, EnergyPer1, EnergyPer2, EnergyPer3(EnergyHalfPoint), EnergyPer4, and EnergyPer5 for boundary effects.

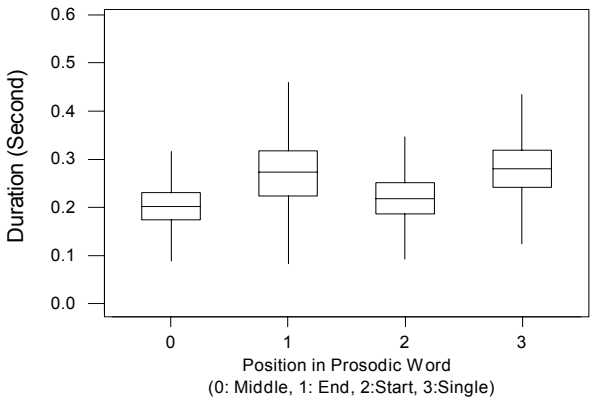


Figure 5.9 Boxplots of Duration by boundary type

According to the position of a syllable in a prosodic word, syllables can be classified into 4 categories, which are, single syllable prosodic word, start, middle and end of a multiple syllable words.

We draw boxplot for Duration, EnergyStart, EnergyHalfPoint, and EnergyEnd as shown in Figure 5.9, 5.10, 5.11, and 5.12 respectively. Each of the figures shows that there are different patterns for different boundary syllable types. This shows that these parameters can make more or less distinction between boundary types. Examination of the rest of parameters gives similar conclusion. Generally, the parameters provide some distinctions for different types of unit (in terms of position in prosodic word).

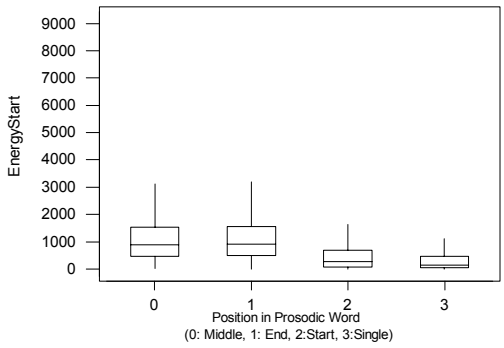


Figure 5.10 Boxplots of EnergyStart by boundary type

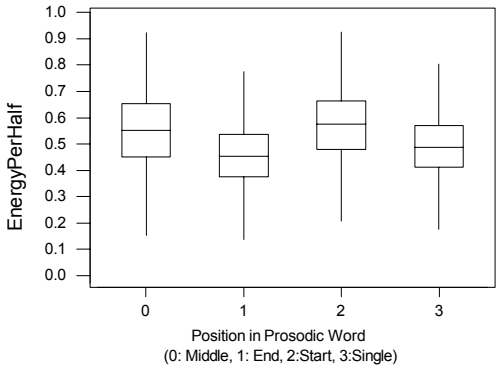


Figure 5.11 Boxplots of EnergyHalfPoint by boundary type

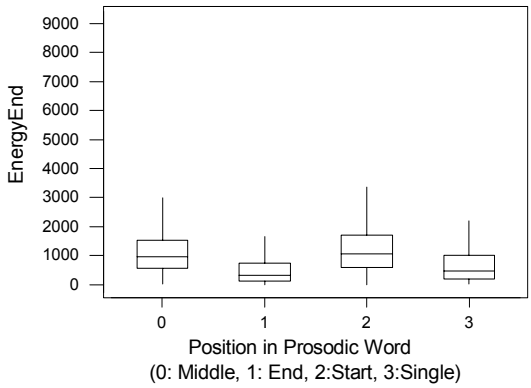


Figure 5.12 Boxplots of EnergyEnd by boundary type

Like what we have done for tone, we also investigate the parameters from recognition view. We investigate the accuracy of predicting the end of prosodic word (EndOfPW) only. The reason is that end syllable (EndOfPW) and start syllable (StartOfPW) of prosodic word always appear as neighbors. CART approach is used for the recognition. The features for this recognition are as the following:

- Duration and Energy (Max value, Sum value and RMS value)
- PitchMean, PitchRange
- EnergyPer1, EnergyPer2, EnergyPer3, EnergyPer4, EnergyPer5

Training Data		
Actual Class	Total Cases	Percent Correct
0	18,373	86.4
1	13,698	76.1
Testing Data		
Actual Class	Total Cases	Percent Correct
0	18,273	82.5
1	13,698	72.5

Table 5.3 Recognition result of StartOfPW

The recognition result of EndOfPW is shown in Table 5.3. It shows that the accuracy of end syllable of prosodic word is 76.1% for training data and 72.5% for testing data. It shows that the above parameters can help to distinguish boundaries. We should note the following factors that are affecting accuracy as well:

- Some of the prosodic word break cannot be recognized correctly because the breaks are not clear when the speaker read the utterance. In real speech, sometimes, there is no clear distinction whether a word break is a prosodic word boundary or not. This is observed in our speech corpus.

- Syllable identity is not included in recognition. Therefore, we missed some discriminating factors in recognition. The reason to exclude syllable identity is that we want to exclude the effect of text information, which is contained in syllable identity. Some of breaks can be identified from syllable identity. For example, “DE5” is the pronunciation of character “的”, which is usually an end syllable of prosodic word.

- Boundary is placed between two syllables. Therefore, boundary effect is a combined effect of two syllables. This obtained accuracy is only obtained from the syllable before the break.

We calculate the total accuracy and find the total accuracy for break is 82.0% for training data and 78.2% for testing data. We should note that if randomly assigning prosodic break types to break between syllables, the accuracy should be 50% in theory.

We conducted a listening test for syllables. Each listening is to judge whether the syllable is an end syllable of prosodic word. 3 persons listened to 200 syllables and achieved an accuracy of 72.1%. This result is even worse than that by break recognition. The reason for this result is that break is prominent only when multiple syllables are placed together, and many of the breaks between syllables sound between break and non-break. The result shows that our recognition rate is sufficiently good. Hence, the parameters help to describe break type.

	Duration	Energy Start	Energy End	Energy Max	Energy RMS	Energy Sum	Energy Per1	Energy Per2	Energy Per3	Energy Per4
EnergyStart	0.022									
EnergyEnd	-0.431	-0.075								
EnergyMax	0.041	-0.043	0.180							
EnergyRMS	-0.109	0.019	0.294	0.950						
EnergySum	0.181	0.004	0.139	0.954	0.948					
EnergyPer1	-0.071	-0.403	0.391	0.013	-0.064	-0.079				
EnergyPer2	-0.125	-0.317	0.455	-0.069	-0.102	-0.135	0.939			
EnergyPer3	-0.174	-0.262	0.506	-0.147	-0.130	-0.179	0.856	0.963		
EnergyPer4	-0.217	-0.217	0.553	-0.217	-0.152	-0.215	0.765	0.888	0.964	
EnergyPer5	-0.258	-0.168	0.603	-0.273	-0.169	-0.247	0.652	0.774	0.866	0.946

Table 5.4 Correlation values between break related variables

We next examine the relationship between the parameters. We calculate the correlation values between the parameters. The values are as listed in Table 5.4. From the table, we see that Duration has low correlation values with other parameters. EnegyStart has low correlation values with others. EnergyEnd has moderate correlation values with EnergyPer1 to EnergyPer5. EnergyRMS, EnergySum, and EnergyMax have high correlation values between each other. EnergyPer1 to EnergyPer5 have high correlation values between each other. Therefore, there is redundancy in the defined parameters.

(3) Parameter Selection

Since there is redundancy in our candidate parameters, in this part, we conduct experiments to select representative parameters from the candidate parameter set. Using clustering approach, we select parameters that have less correlation values between each other. The procedure of clustering is an agglomerative hierarchical method that begins with all parameters separate, each forming its own cluster. In the first step, the two parameters closest together are joined. In the next step, either a third parameter joins the first two, or two other parameters join into a different cluster. This process will continue until all clusters are joined into one. At last, we need to decide the number of clusters.

The clustering process can be shown as in a dendrogram as shown in Figure 5.13. Figure 5.14 shows that similarity levels at each step of clustering. The similarity, s(i,j), between two clusters i and j is given by:

$$s(i,j) = 100(1-D(i,j)) \qquad (5.18)$$

where D(i,j) is the distance between two clusters. In the figure, axis x is the number of step. Axis y means, at this step, the parameters have similarities above this value have been combined. In the figure, we can see that there is an abrupt change from similarity 81.4 to 65.7 at step 13. Therefore, we cut the dendrogram at similarity level 80.

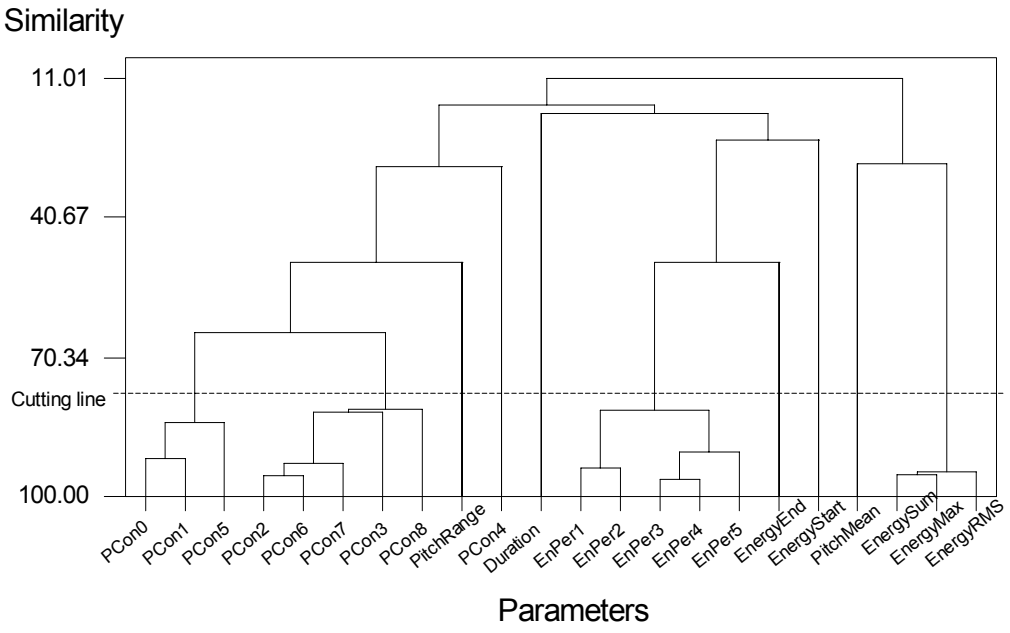


Figure 5.13 Dendrogram for clustering parameters

Drawing a cutting line on the dendrogram at similarity value 80 in Figure 5.13, we get the final clusters. The final clusters are shown in Table 5.5. The table shows the parameters in each cluster. We select one parameter from each cluster as a

representative of the cluster. The third column is the parameters we finally determined in TTS system.

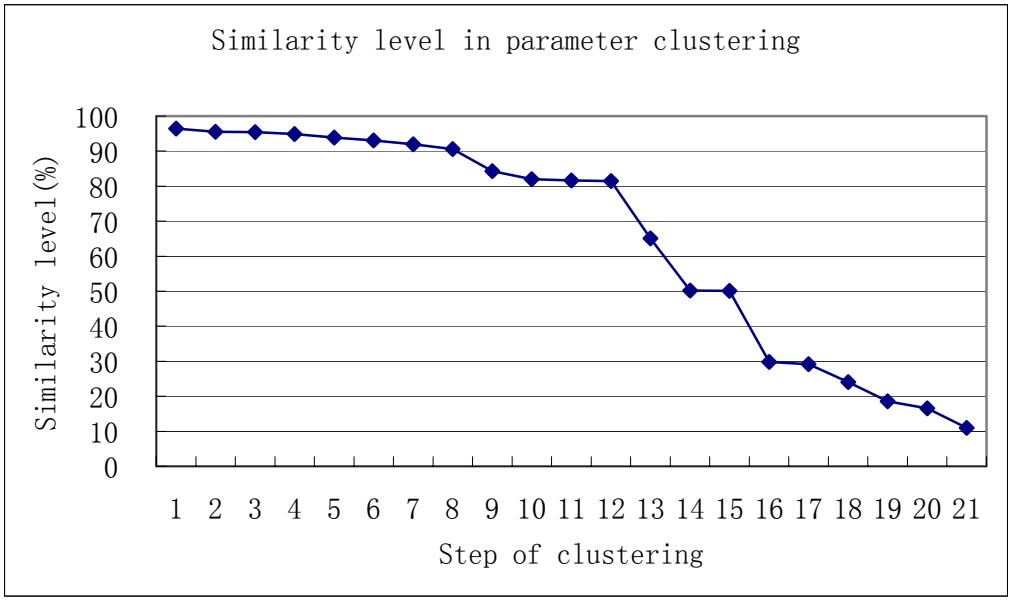


Figure 5.14 Similarity level in paramter clustering step

In the table, we see that PCon0 , PCon1 and PCon5 fall in one cluster. We choose Pcon0 (PitchStart) because it is the first value in the contour. Accurately determining this value will help to maintain the prosody smoothness between this syllable and previous syllable in utterance. Pcon4 constitutes a cluster itself. It is coincident that the value is actually the pitch value at the middle point of the contour. PCon2, Pcon3, Pcon6, Pcon7, and Pcon8 belong to one category. We choose Pcon8 (PitchEnd) as representative of this cluster. Selecting this parameter has the same reason as selecting Pcon0 in cluster 3 for the purpose to maintain continuous in pitch between two syllables.

We also see that the three types of energy values fall into 1 cluster. We select the RMS energy as their representative, as this is a preferred value as we described in 5.3.1.

Parameters EnPer1, EnPer2, EnPer3, EnPer4 and EnPer5 are clustered together. We select the middle value EnPer3 (EnergyHalfPoint) as representative.

Cluster No.	Parameters in the cluster	Selected Parameter
1	Duration	Duration
2	PitchMean	PitchMean
3	PCon0 PCon1 PCon5	Pcon0
4	PCon2 PCon3 PCon6 PCon7 PCon8	Pcon8
5	PCon4	Pcon4
6	EnergySum EnergyMax EnergyRMS	EnergyRMS
7	EnPer1 EnPer2 EnPer3 EnPer4 EnPer5	EnergyHalfPoint
8	PitchRange	Pitchrange
9	EnergyStart	EnergyStart
10	EnergyEnd	EnergyEnd

Table 5.5 Final clusters in parameter clustering

	Duration	Pitch Mean	Pitch Con0	Pitch Con4	Pitch Con8	Pitch Range	Energy Start	Energy End	Energy RMS
PitchMean	-0.219								
PitchCon0	0.112	-0.217							
PitchCon4	-0.122	0.191	-0.378						
PitchCon8	-0.086	0.013	-0.572	-0.416					
PitchRange	0.171	-0.105	0.459	0.033	-0.523				
EnergyStart	0.022	-0.122	0.184	-0.087	-0.079	0.09			
EnergyEnd	-0.431	0.370	-0.198	0.016	0.235	-0.154	-0.075		
EnergyRMS	-0.109	0.328	-0.004	-0.006	-0.037	0.127	0.019	0.294	
EnergyPerHalf	-0.174	0.245	-0.235	0.087	0.213	-0.244	-0.262	0.506	-0.130

Table 5.6 Correlation values between selected parameters

We examine the correlations between the selected parameters. The correlations are shown in Table 5.6. We see from the table that the highest correlation in absolute value is 0.572. Most correlation values are very low. Therefore, the selected parameters have little redundancy as we expected. Models for predicting the 10 parameters will be built later in this chapter. These parameters will be used in unit selection process in Chapter 6.

(4) Summary of Parameter Determination

In this work, we proposed an approach to evaluate and select parameters for unit selection based synthesis. We summarize the steps for parameter determination as follows:

(a) Define all the candidate parameters.

(b) Evaluate whether the intended parameters are the discriminating parameters for the prosodic effect from statistical view.

(c) Evaluate whether the discriminating parameters are sufficient to describe the prosodic effect using recognition approach.

(d) If the parameters are not sufficient to describe the intended prosodic effect, go to step (a) to define more parameters.

(e) If the parameters are sufficient, perform a parameter clustering process. This step groups parameters together into a tree structure.

(f) Determine the final clusters, and select one parameter from each cluster as representative parameter. This step removes the redundancy and determines a final set of parameters.

Note that this work is only an example for doing similar work. We can identify the following generality for this approach:

Parameters: This work defined a candidate parameter set of 22 parameters in 5.3.3. However, there is no limit of defined parameters. The parameters were defined from 3 aspects of prosody (pitch, duration and energy). However, the defined parameters are not the only choice to do the work. One can certainly define a new set of equivalent parameters to achieve the same goal. Moreover, one can also define parameter beyond pitch, duration and energy. In principle, any acoustic parameters can be defined as long as they are correlates of some perceptual effects.

Perceptual Effects: In this work, we highlight the ability of describing the perceptual prosodic effects, tone and break. However, there is no limit for such prosodic events. The idea can be used in other prosodic events. If there are sufficient labeled data and sufficient parameters, we can also evaluate and find parameters for describing any abstract prosodic events, for example, stress, emotion status (such as happiness, sadness, surprise), etc.

Language: This approach is also not limited to Chinese speech only. It can work for any language. To apply to a new language, a corpus of this language should be built. The parameters suitable for this language should be defined. To generate good prosody in speech, we also need to concentrate on some prosodic effects of this language (such as tone and break in this work).

5.6.2 Single Feature in Prediction

We now examine the discriminating ability of the features in prosody parameter prediction. To make sure the results obtained are reliable, we divided the data into two halves. For each half, we use each feature as prediction feature, then we calculated the correlation values. In the following tables, correlation1 and correlation2 are the values obtained from the two halves of the data. In this following discussion, for consistent results, we will use the average value of correlation1 and correlation2 to explain our findings. (Explanation of the methods for this experiment can be found in Section 5.5.2.)

We choose to predict three parameters (PitchMean, Duration, and Energy) because the parameters are the most important parameters for genreal prosody (see Section 2.2).

(1) Factors Affecting Pitch Mean

Table 5.7 shows correlation of the factors in predicting PitchMean of syllable. Examining the table, the following facts are found:

1. **Most important factor:** The correlation obtained using tone alone is 0.654, which is the highest. Therefore, tone is the most important factor in determination of pitch mean of syllables.

2. **Syllable and neighboring syllables:**

Category	Feature	Correlation1	Correlation2	Average
Current Syllable	CurrIntial	0.185	0.176	0.181
	CurrFinal	0.213	0.091	0.152
	CurrTone	0.662	0.645	0.654
Context Syllables	PrevInitial	0.242	0.243	0.243
	PrevFinal	0.233	0.221	0.227
	PrevTone	0.194	0.200	0.197
	NextInitial	0.269	0.280	0.275
	NextFinal	0.233	0.082	0.158
	NextTone	0.221	0.217	0.219
Current Word	WordPOS	0.228	0.230	0.229
	WordLen	0.054	0.028	0.041
Context Words	PrevWordPOS	0.284	0.282	0.283
	PrevWordLen	0.110	0.131	0.121
	NextWordPOS	0.182	0.198	0.190
	NextWordLen	0.030	0.037	0.034
Location in Word	LocInWord	0.149	0.167	0.158
	WordStart	0.148	0.167	0.158
	WordEnd	0.195	0.195	0.195
Prosodic Word	PWLen	0.046	0.050	0.048
	LocInPW	0.000	0.000	0.000
	PWStart	0.250	0.260	0.255
	PWEnd	0.321	0.328	0.325
Intonation Phrase	IPType	0.023	0.015	0.019
	LocInIP	0.313	0.337	0.325
Break Type	BreakBefore	0.322	0.331	0.327
	BreakAfter	0.363	0.375	0.369

Table 5.7 Comparison of factors determining pitch mean

a. The tone of the syllable (0.654) is important, while the initial and final of the syllable (0.181 and 0.152) are less important in predicting pitch mean.

b. It is interesting to look at the correlations obtained by initial and final of the previous (0.243 and 0.227) and next syllable (0.275 and 0.158). The values are larger than the corresponding values obtained by the current syllable. It shows that the context of a syllable could be more important even than the initial and final type of the syllable itself. This can be explained that in some words, the previous or next syllable of the current syllable is more important in determining the nature of the words.

3. **Word level:**

a. The POS of the current word (0.229) is important, while its length (0.041) is less important. This shows that pitch mean is more determined by the syntactical property (e.g. POS) rather than the form (e.g. length) of word.

b. The POS types of previous word (0.283) and next word (0.190) also have larger impact on the pitch mean than lengths (0.121 and 0.034).

4. Word and prosodic word:

a. Length of word (0.041) and length of prosodic word (0.048) are less important in determining the pitch mean compared with other factors (e.g. POS, Start and End).

b. Start and end of prosodic word (0.255 and 0.325) have bigger effect on pitch mean than start and end of word (0.158 and 0.195). This shows that prosodic word is more meaningful in predicting pitch mean.

5. **Intonation phrase:**

a. Intonation type (IPType, or Major phrase type) (0.019) is less important in pitch mean prediction. The reason is that intonation type normally affects the syllables in the final part of the utterance, which are only a very small part of all syllables in the corpus.

b. The location of syllable in phrase (0.323) is an important input value. This can be explained by the fact that the general pitch contour has a trend of going down in an utterance.

6. **Break types:**

a. Break types before and after a syllable (0.327 and 0.369) is very important in predicting pitch mean.

b. Note that prosodic word breaks are major parts of break types. Comparing values of the start and end of prosodic word (0.255 and 0.325) with break types (0.327 and 0.369), we find prosodic word break take an important part in break types for predicting pitch mean.

7. Conclusion:

In summary, we find that, in determining pitch mean:

a. Current tone is the greatest factor.

b. Surrounding syllables have a big impact.

c. POS of word is more important than length of word.

d. Prosodic word is more meaningful than word.

e. Length of prosodic word is less important than the start and end positions of prosodic word.

f. Breaks before and after a syllable have great impacts.

g. Location in phrase is more important than type of intonation phrase.

(2) Factors Affecting Duration

A comparison of factors determining duration is listed in the Table 5.8. From the table, we have the following findings:

1. **Most important factor:** Break type after a syllable (0.438) has the largest value in determining duration.

2. **Syllable and neighboring syllables:**

Category	Feature	Correlation 1	Correlation 2	Average
Current Syllable	CurrIntial	0.343	0.330	0.337
	CurrFinal	0.248	0.132	0.190
	CurrTone	0.180	0.181	0.181
Context Syllables	PrevInitial	0.088	0.081	0.085
	PrevFinal	0.092	0.098	0.095
	PrevTone	0.031	0.033	0.032
	NextInitial	0.312	0.315	0.314
	NextFinal	0.261	0.080	0.171
	NextTone	0.220	0.218	0.219
Current Word	WordPOS	0.216	0.226	0.221
	WordLen	0.088	0.089	0.089
Context Words	PrevWordPOS	0.101	0.087	0.094
	PrevWordLen	0.033	0.025	0.029
	NextWordPOS	0.228	0.240	0.234
	NextWordLen	0.054	0.050	0.052
Location in Word	LocInWord	0.108	0.118	0.113
	WordStart	0.102	0.113	0.108
	WordEnd	0.193	0.216	0.205
Prosodic Word	PWLen	0.217	0.222	0.220
	LocInPW	0.000	0.000	0.000
	PWStart	0.124	0.137	0.131
	PWEnd	0.412	0.430	0.421
Intonation Phrase	IPType	0.040	0.033	0.037
	LocInIP	0.048	0.053	0.051
Break Type	BreakBefore	0.133	0.149	0.141
	BreakAfter	0.428	0.447	0.438

Table 5.8 Comparison of factors determining duration

a. The initial, final, and tone of current syllable (0.337, 0.190 and 0.181) have great effects for duration of syllable. Among them, initial is the most important.

b. The values for initial, final, and tone of next syllable (0.314, 0.171 and 0.219) are very high. This shows that the next syllable has great influence for the duration of the syllable. This can be explained that, when uttering the current syllable, a speaker will get ready for uttering the next syllable. For different following syllables, a speaker will take different amount of time to adjust speech organ.

c. On the other hand, the values for initial, final, and tone of previous syllable (0.085, 0.095 and 0.032) are very low. This shows that the previous syllable gives little contribution for the duration of current syllable.

3. **Word level:**

   a. Similar to pitch mean and pitch range, POS of word (0.221) is important than length of word (0.089).

   b. POS of next word (0.234) is more important than POS the previous word (0.094).

4. **Word and prosodic word:**

   a. Length of prosodic word (0.220) has a significant effect on duration, while length of word (0.089) does not.

   b. End of prosodic word (0.421) has more influence on duration that end of word (0.205). This means duration is sensitive for the last syllable of a prosodic word.

5. **Intonation phrase:** Intonation type (0.037) and location of the syllable in phrase (0.051) have no significant effect on duration.

6. **Break types:** Break types after the current syllable (0.438) is much more important than break types before the current syllable (0.141) in determining duration.

7. **Conclusion:**

   In summary, we have the following findings:

   a. The most important factor for duration is the break type after the syllable. The break type before the syllable is less important.

   b. The duration of a syllable is more determined by the next syllable than the previous syllable.

   c. Prosodic word is more meaningful in determining duration than word.

   d. POS of word is important, while length of word is not.

   e. POS of next word is more important than POS of previous word.

(3) Factors Affecting Energy

Table 5.9 lists correlation values of energy (EnergyRMS) obtained by all the features. We found that:

1. **Most important factor:** The final of the current syllable (0.370) has the greatest influence on energy.

2. **Syllable and neighboring syllables:** Initial, final and tone of previous syllable (0.250, 0.229 and 0.156) have a larger influence on energy than those of the next syllable (0.167, 0.102 and 0.123).

3. **Word level:**

   a. POS of word (0.161) and length of word (0.102) have a moderate effect on energy.

   b. POS of previous word (0.283) and POS of next word (0.140) are more important in determining energy than length of the words (PrevWordLen: 0.089, NexWordLen: 0.056).

c. POS of previous word (0.283) is more important than POS of next word (0.140). This finding is consistent with that previous comparison on syllable level.

4. **Prosodic word:** Start of prosodic word (0.115) is more important than end of prosodic word (0.015).

Category	Feature	Correlation 1	Correlation 2	Average
Current Syllable	CurrIntial	0.312	0.333	0.323
	CurrFinal	0.485	0.254	0.370
	CurrTone	0.170	0.186	0.178
Context Syllables	PrevInitial	0.243	0.257	0.250
	PrevFinal	0.223	0.235	0.229
	PrevTone	0.145	0.166	0.156
	NextInitial	0.172	0.162	0.167
	NextFinal	0.142	0.062	0.102
	NextTone	0.127	0.119	0.123
Current Word	WordPOS	0.155	0.166	0.161
	WordLen	0.105	0.099	0.102
Context Words	PrevWordPOS	0.273	0.292	0.283
	PrevWordLen	0.079	0.099	0.089
	NextWordPOS	0.130	0.150	0.140
	NextWordLen	0.064	0.047	0.056
Location in Word	LocInWord	0.092	0.088	0.090
	WordStart	0.090	0.088	0.089
	WordEnd	0.038	0.027	0.033
Prosodic Word	PWLen	0.110	0.076	0.093
	LocInPW	0.000	0.000	0.000
	PWStart	0.115	0.114	0.115
	PWEnd	0.000	0.030	0.015
Intonation Phrase	IPType	0.000	0.000	0.000
	LocInIP	0.318	0.327	0.323
Break Type	BreakBefore	0.277	0.291	0.284
	BreakAfter	0.160	0.147	0.154

Table 5.9 Comparison of factors determining Energy

5. **Intonational phrase:** Location in intonational phrase (0.323) has a better discriminating ability than type of the phrase (0.030).

6. **Break types:** Break before the syllable (0.284) is much more important than break after the syllable in determining energy (0.154)

7. **Conclusion:**

    a. The greatest factor in determining energy is the final of the syllable.

    b. Syllable before the current syllable has a better discriminating ability in determining energy than that after the current syllable.

    c. Break type before the current syllable is more important in determining energy than that after the current syllable.

(4) Summary of the Analysis

We have the following findings from previous analysis:

- PitchMean is mostly determined by tone; Duration is mostly determined by break type after the syllable; Energy is mostly determined by final of the syllable.

- PitchMean is affected by both previous and next syllable; Duration is more affected by next syllable; Energy is more affected by previous syllable.

- POS of word is more important than length of word in predicting predict PitchMean.

- Prosodic word is more meaningful than word in predicting the parameters.

- Breaks before and after syllable are equally important in determining PitchMean; Break after syllable are more important in determining Duration; Break before syllable are more important in determining Energy.

- Location of syllable in utterance greatly affects PitchMean and Energy. However, it has little effect on duration.

5.6.3 Combined Features for Prediction

In this part, we examine the prediction ability of combined features. This is done by using stepwise training of regression tree. (Explanations of methods for this experiment can be found in Section 5.5.2.) Among the 10 parameters we determined in 5.6.2, We will examine the following parameters: PitchMean, Duration, and EnergyRMS. The reason to examine them is that they are parameters to describe the general property of prosody.

Step	Feature	Correlation achieved
1	CurrTone	0.6490
2	BreakBefore	0.7536
3	BreakAfter	0.8029
4	LocInIP	0.8340
5	PrevTone	0.8524
6	NextTone	0.8617
7	PWLen	0.8668
8	WordPOS	0.8709
9	CurrInit	0.8757
10	PrevPOS	0.8778
11	CurrFinal	0.8787
12	NextWordLen	0.8796
13	PrevWordLen	0.8800
14	NextInit	0.8803
15	PrevInit	0.8805
16	NextPOS	0.8807
17	LocInPW	0.8810
18	NextFinal	0.8811
19	EndOfPW	0.8811

Table 5.10 Stepwise training for PitchMean

(1) Stepwise Training of PitchMean

There result of stepwise training of regressing tree is shown in Table 5.10 and Figure 5.15. The correlation value obtained by adding each feature is shown in the table. The features are listed in descending order according to its importance in the prediction.

We can see from the figure that the achieved value changes quickly in the first five steps. Therefore, the first a few features have the greatest contribution in predicting pitch mean. In the table, we can see the most important features are:

- Tone of the syllable
- Break type before the syllable
- Break type after the syllable
- The location of the syllable in intonational phrase
- Tone of the previous syllable
- Tone of the next syllable

All the above facts show us that:

- PitchMean is one of the discriminating parameters for tone
- PitchMean changes at boundary syllables (sensitive to breaks before and after the syllable)
- PitchMean is greatly determined by tones of the current syllable and surrounding syllables

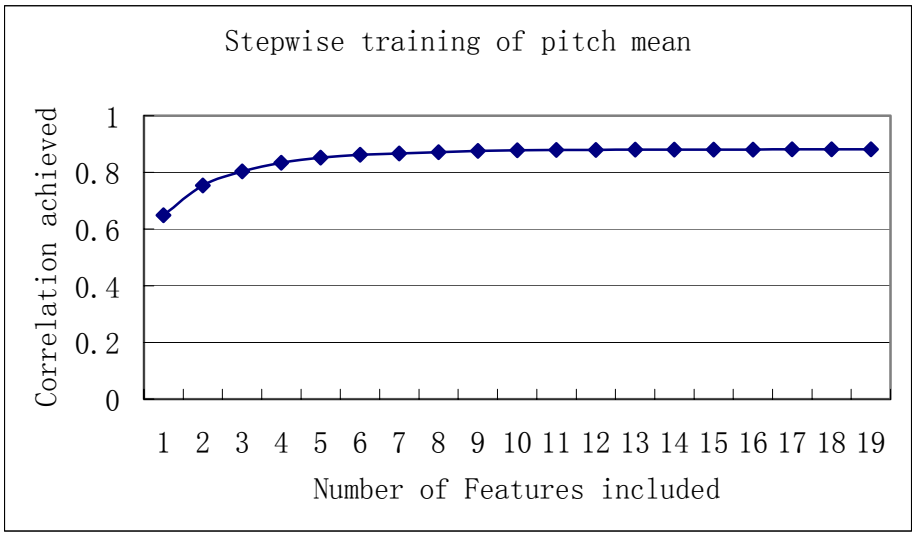


Figure 5.15 Stepwise training of PitchMean

(2) Stepwise Training of Duration

The result for Duration is shown in Table 5.11 and Figure 5.16. In the figure, we can see that the achieved correlation value becomes stable after six steps. The most important factors are:

- Break after the current syllable

- Initial of the current syllable
- Final of current syllable
- Tone of the current syllable
- POS types of the next word
- Break type before the current syllable

The facts show:

- Break is the most important factor for duration. Therefore, this parameter is discriminating factor for boundary (break).
- Syllable identity with tone is the second important factor for duration.
- POS type of the word after the syllable is an important factor.

Step	Feature	Correlation achieved
1	BreakAfter	0.4717
2	CurrInitial	0.6261
3	CurrFinal	0.6947
4	CurrTone	0.7267
5	NextWordPOS	0.7421
6	BreakBefore	0.7501
7	WordPos	0.756
8	LocInIP	0.762
9	PWLength	0.7656
10	NextWordLen	0.7686
11	NextTone	0.7709
12	NextInitial	0.7728
13	PrevTone	0.774
14	PrevInitial	0.7745
15	PrevWordLen	0.7748

Table 5.11 Stepwise training for Duration

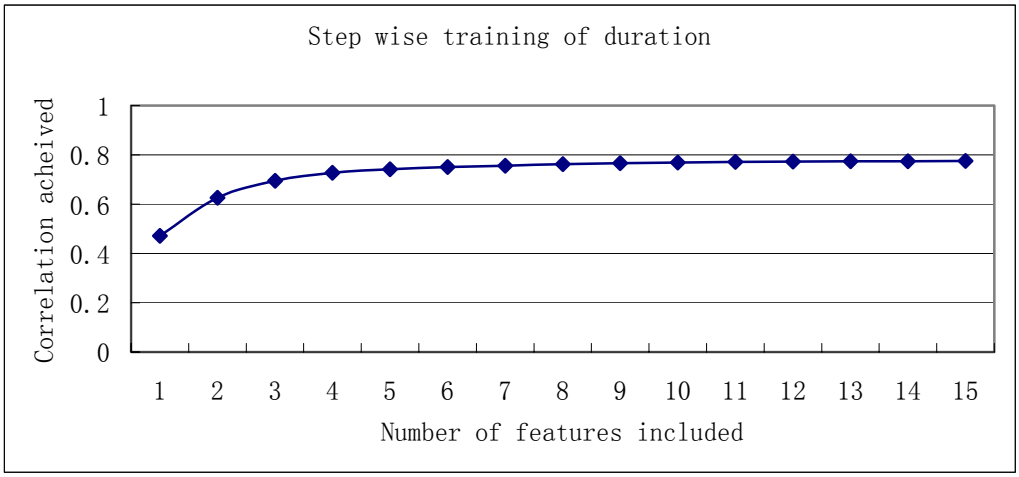


Figure 5.16 Stepwise training of Duration

Order of feature	Feature	Correlation achieved
1	CurrFinal	0.5184
2	PrevInitial	0.6540
3	CurrInitial	0.6980
4	CurrTone	0.7223
5	LocInIP	0.7485
6	BreakAfter	0.7586
7	PrevTone	0.7608
8	PWLen	0.7629
9	WordLen	0.7640
10	PrevWordLen	0.7652
11	WordPOS	0.7660
12	NextWordLen	0.7673
13	NextTone	0.7680
14	NextWordPOS	0.7687
15	LocInWord	0.7690

Table 5.12 Stepwise training for Energy

(3) Stepwise Training of EnergyRMS

Table 5.12 and Figure 5.17 show the result of stepwise training for EnergyRMS of a syllable. The value of achieved correlation increases quickly in the first six steps. The first six features are most important for the prediction of EnergyRMS. The most important features are:

- Final of the current syllable
- Initial of the previous syllable
- Initial of the current syllable
- Tone of the current syllable
- Location of the syllable in intonational phrase
- The break type after the syllable

The facts show that:

- Energy is mostly dependent on the final of the current syllable.
- Syllable identity is the main factor for the parameter.
- Location of the syllable in intonational phrase is an important factor. The reason is that energy has a downtrend from the start to end of an intonational phrase (most of time, intonational phrase is an utterance.).

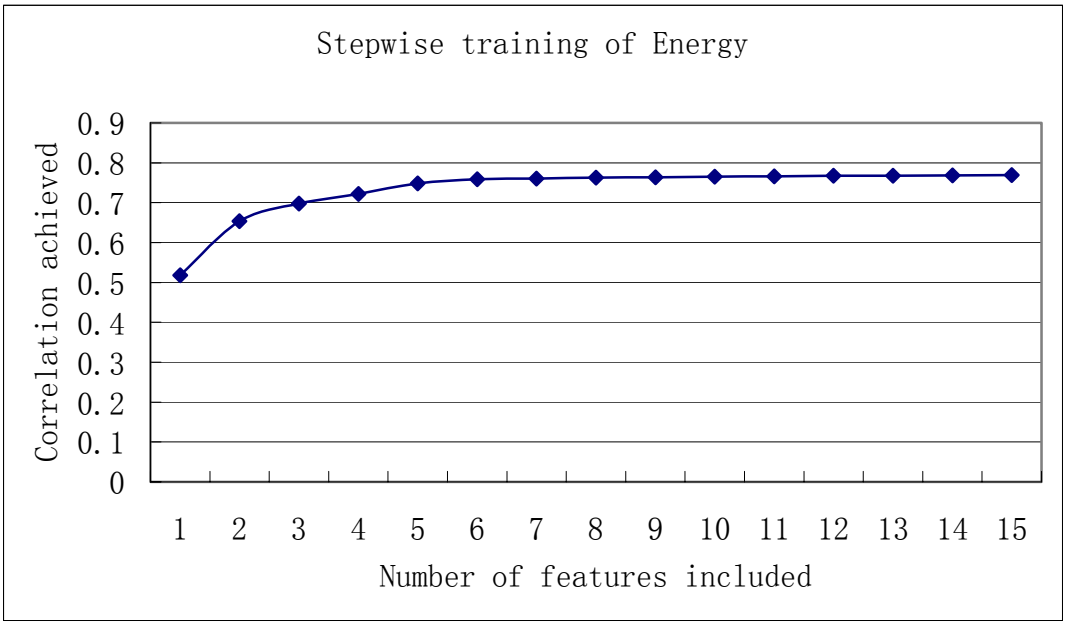


Figure 5.17 Stepwise training of Energy

(4) Summary of the Analysis

For all the stepwise training above, we can see that the most influential input features for prosody prediction are:

- Initial, final and tone of the current syllable
- Initial, final and tone of the previous and next syllables
- The break types before and after the current syllable
- Location of the syllable in the intonational phrase

Examining features for PitchMean, we find that the most important factor is tone of the syllable. We also find that tone of previous syllable, tone of next syllable, breaks around the syllable, and location of the syllable in utterance play important roles. However, the final of the syllable, which is the actual carrier of the tone, is not an important factor in predicting PitchMean. That means PitchMean of a tone contour is almost independent of the sound that carries the tone.

We examine the parameter Duration and find that, besides break types before and after the syllable, syllable identity (the initial, final and tone of the syllable) is an important factor for the prediction. The reason why syllable identity is important is that different syllables have different intrinsic durations.

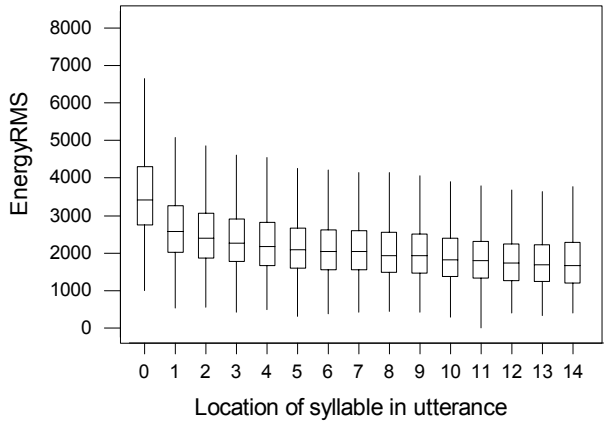


Figure 5.18 EnergyRMS changing with location of syllable in utterance.

Examining EnergyRMS, we find that EnergyRMS is determined by final of the syllable mostly. The initial and tone of the syllable are in the third and forth position. We find that location of the syllable in intonational phrase is one of the important factors for prediction. This can be confirmed by Figure 5.18. This is a boxplot for Energy, classified by location of syllable in utterance. The boxplot of EnergyRMS figure shows that the EnergyRMS has a decreasing trend with the change of location in utterance.

5.6.4 Prediction of All Parameters

The prosody parameters are predicted using CART. In this experiment, we first randomize the order of the data items in the data set. Then we divide the data set into training set and testing set, which include 80% and 20% of the data items respectively. This experiment is conducted without using stepwise training because stepwise training is extremely slow. The minimal node size is set to 20. The results are shown in Table 5.13. Here we list the Root Mean Squared Errors (RMSE) and correlation values of the predicted parameters.

In the table, we can see that the PitchMean has the highest correlation value (0.8791 for training data and 0.8526 for testing data) among all the parameters.

PitchStart, PitchEnd, EnergyEnd, EnergyHalfPoint are parameters in the second highest correlation value group. This shows these parameters are relatively more stable than others are.

	Training data		Testing data	
Parameter	RMSE	Correlation	RMSE	Correlation
PitchMean	25.32 Hz	0.8791	28.85 Hz	0.8526
PitchStart	24.85 Hz	0.7753	27.35 Hz	0.7337
PitchEnd	25.96 Hz	0.7773	27.97 Hz	0.7512
PitchMiddle	8.70 Hz	0.6552	9.66 Hz	0.6049
PitchRange	31.44 Hz	0.6771	34.56 Hz	0.5982
Duration	0.037 Sec	0.7262	0.040 Sec	0.6723
Energy	447.3	0.7346	621.3	0.6614
EnergyStart	521.78	0.7382	576.58	0.6910
EnergyHalfPoint	0.083	0.7961	0.091	0.7486
EnergyEnd	490.00	0.7598	534.10	0.7207

Table 5.13 Result of the prosody parameter prediction

The lowest correlation value obtained is for PitchRange (0.6771 for training data and 0.5982 for testing data). Duration, EnergyRMS and EnergyStart have relatively low correlation values. This shows that these three parameters are not so stable. Pitch range can change with stress degree of a syllable, which cannot be easily derived from text input, and is not included in the features for prediction. Therefore, the accuracy of PitchRange is relatively low. Duration is related to breaks between syllables. However, the time length of a break is flexible. Therefore, accuracy of Duration is relatively low. Energy is determined by volume of speech. It is possible that the volume levels vary among different utterances. Therefore, Energy has a relatively low accuracy in prediction.

The accuracies cannot be easily compared with those of other research work. The reasons are: (1) The definition of the parameters and the corpus used are different. (2) The accuracies of parameters are not the only measures to evaluate the parameters.

The significance of the parameters is that they are intended to describe some perceptual effects. The selection process of the parameters shows that the parameters capture the information of the perceptual effects. Another difference of the work from

the other approaches is that a standard deviation is also predicted, which will measure the variation of the parameters.

No matter how well the parameters are defined or predicted. Its effectiveness can only be shown when the prosody is applied to real TTS process. We will apply the generated prosody to unit selection-based synthesis approach in Chapter 6. The synthetic speech will be evaluated in Chapter 7.

5.7 Summary

This chapter describes the process of design, evaluation and determination of the prosody parameters. First, I introduce the prosody parameters and review the prosody prediction approaches. Second, the problem of prosody parameters for unit selection is stated. The solutions to the problems are proposed. Third, the parameters are defined. The processes for evaluating and selecting parameters are described. A clustering approach is adopted to determine the final parameter set. Finally, relationships between parameters and features are investigated.

In this chapter, I proposed an approach to determining parametric prosody representation for unit selection based synthesis. This approach solved the following problems that encountered in unit selection based speech synthesis. (1) The approaches for evaluating prosody parameters have been given. This helps to determine whether the parameters are sufficient to describe perceptual prosody effects (e.g. tone and break). (2) The approach for determining final parameter set has been given. The approach can determine a parameter set, which is concise but sufficient. (3) Using a regression tree approach, the prosody models predict the prosodic parameter as well as the standard deviation of the class to which it belongs. This makes it possible to measure mismatch in unit selection based synthesis.

This work provides a solution for determining a set of prosody parameters suitable for unit selection based synthesis. The selected parameters describe not only the general prosody of speech but also the important perceptual prosody effects. The proposed approach can be extended to languages other than Chinese, or to prosody properties other than break and tone.

For the prosody description for Chinese, I discovered that energy contour (or its equivalent) helps to describe boundary units. I discovered the relationship between the prosody parameters and the features for prediction. This result helps to understand the prosody parameters and features better. This is useful when building prosody models of different sizes, in which some factors can be neglected.

Chapter 6 Unit Selection with Prosody

In this chapter, we describe how the prosody parameters that are determined in the previous chapter are integrated into the unit selection process. First, an introduction to speech synthesis techniques is given. Then, we describe the corpus-based unit selection approach. Next, we define the cost function, into which the prosody parameters are integrated. Finally, the weights for the subcosts are determined.

6.1 Introduction

The strategies of synthesizing speech on computer can be classified into three major categories (Flanagan, 1972), which are articulatory synthesis, formant synthesis, and concatenation synthesis. Articulatory synthesis attempts to model the human speech production systems, while formant synthesis and concatenation synthesis attempt to only model resultant speech. Formant synthesis generates speech with the support of a database of rules. Concatenation synthesis works with a database of pre-recorded speech pieces. Unit selection based approach belongs to the category of concatenation synthesis.

6.1.1 Unit Selection-Based Synthesis

(1) Unit Selection-Based Concatenation Synthesis

Normal concatenation synthesis works by keeping a small unit inventory during synthesis. A unit is selected and then modified using signal processing techniques according to prosody features. Synthesis by this way can generate speech with relatively high quality. However, the synthetic speech is more or less distorted due to the signal processing process.

A simple idea of generating good speech is to store large quantities of speech segments of human speech in a database and, when generating, concatenate all the needed speech segments together without any modification. Of course the longer the

stored segments selected for the concatenation, the more natural the generated speech. As each speech unit may have many variants in different contexts or prosodic situations, this approach needs a large memory to store a large number of speech segments. The approach was not practical some years ago because of the limitation of computer power and memory. With the development of hardware, the use of large speech corpus as synthetic units for direct concatenation is possible.

This idea was first proposed to minimize the unnaturalness that caused by the concatenation of small synthesis unit inventory. A non-uniform unit concatenation method was proposed by Sagisaka (1988, 1990). The approach eventually developed to the problem of unit selection (Black and Campbell 1995, Hunt and Black 1996). The key idea of unit selection is to select from corpus the longest available strings of phonetic segments that match a sequence of target speech sounds in the utterance to be synthesized, thereby minimizing the number of concatenations and reducing the need for signal processing. The underlying assumption of the unit concatenation synthesis is that the listener will tolerate the occasional spectral and prosodic mismatch in an utterance if the general quality of the speech is similar to natural speech (Mobius, 2000).

Although there are more or less prosody considerations, the use of prosody for unit selection process is weak. Usually, only basic prosody parameters are defined. The parameters are not enough to describe some important prosody properties. (E.g. break). In addition, the variations of prosody parameters are not carefully considered.

(2) Unit Selection-Based Synthesis for Chinese

Unit selection-based speech synthesis (or corpus-based synthesis) has been applied in English and other languages for some years. In recent years, some attempts (Liu, and Wang, 1998; Chu et al. 2001; Wang et al., 2000, Li et al, 2001) have been made in Chinese TTS using unit selection approach in synthesis process. A representative of the existing unit-selection based system is (Chu et al, 2001). The system used a two-step synthesis framework, in which, there is no prosody model. Prosody is assumed to be implicitly contained in text information. In the unit selection process, when selecting a syllable, the cost function considers the unit, its context, and the position

of the unit in a prosodic word (start, middle or end of a prosodic word.) This approach works relatively well with a huge speech inventory. However, the shortcoming of the approach is that it only takes into consideration part of the many factors that affect prosody. Therefore, the selected unit may not a prosodically best one. Hence, the generated speech sometimes may have bad prosody because the selected units do not suit the context.

The use of prosody parameters in cost function to select the best units has been applied for selecting units in a small unit inventory. Wu et al. (2001) proposed a scheme to select phonetically, linguistically best units and then apply prosodic modifications. Prosody is first generated from some stored template using cost functions. Then synthesis units are selected using cost functions, in which prosody is used, and a PSOLA synthesis part is applied to modify prosody. The scheme is useful in a unit selection-based synthesis. However, their prosody model determines prosody parameters from stored templates, in which only limited prosodic factors are considered.

The biggest problem of the unit selection based approaches is that they do not have a good prosody consideration. This limits the quality of the generated speech.

(3) Unit Selection Model

A unit selection model has a well-organized unit database. The database contains the speech units from a large corpus, which is carefully designed to have a good coverage of all phonetic and prosodic variants of each unit. In the database, each speech unit has a number of possible variants, which are suitable to appear in different phonetic and prosodic environments. The large speech corpus is analyzed offline and all the calculated features are stored in a unit database. In the database, each instance of a unit is described by a vector of features. Each feature may be a discrete or continuous value. The features include features of the unit itself and the context of the unit. The features of the unit itself are used for selecting the correct unit that meets the segmental requirement, while the features of context are used for selecting the contextually best unit, which may minimize the discontinuity between the selected units.

The corpus-based concatenation synthesis is actually a pattern matching process. During the synthesis, the work need to do is to select the best units that phonetically and prosodically best match the target units. Meanwhile, the discontinuity between selected units should be kept as small as possible. To meet these requirements, two costs should be defined in synthesis. One is unit cost, which describes how close a selected unit to the desired unit. The other is connection cost, which describes the degree of continuity between the selected units. The whole cost is the weighted sum of the two costs.

(4) Unit Selection Process

The speech synthesis part accepts information from prosody generation part, retrieves the speech unit database to find a proper template for every target speech unit. During the selection process, the phonetic and prosodic constraints will be applied. The smoothness of the concatenation will also be concerned.

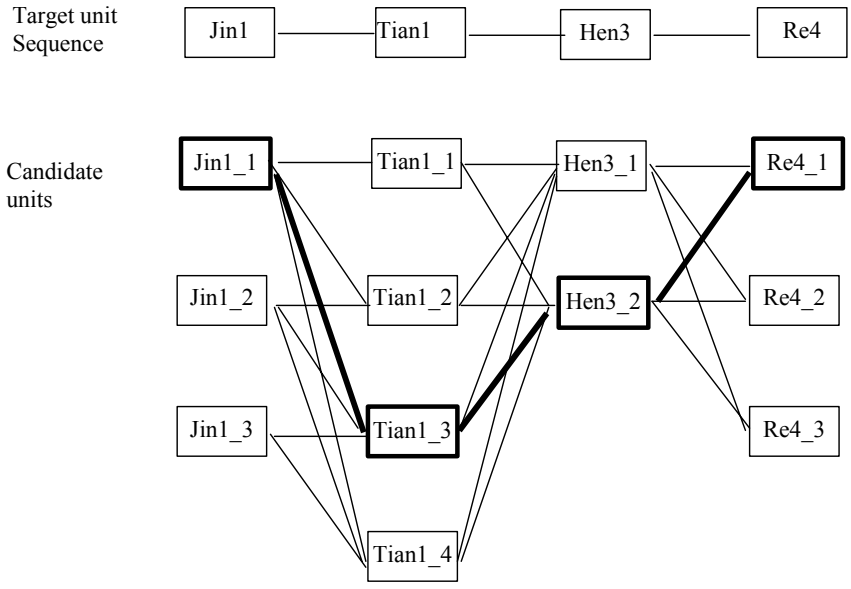


Figure 6.1 Illustration of unit selection

The unit selection process can be illustrated as Figure 6.1. In the figure, the target sentence is “今天很热 (it is hot today)”, which consists of 4 syllables. Each syllable has a set of candidate units. The thick line and thick edge box indicate the

selected unit sequence. In unit selection process, to get the best speech, we have to consider (1) the properness of the unit to target unit, (2) the smoothness between the selected units to be connected. Therefore, the selection process is to find a best path among all the possible paths in the connection lattice. The search process of the path is guided by a cost function, which describes the degree of properness of a unit and degree of smoothness between two units.

6.1.2 Problems of Prosody in Unit Selection

The performance of unit selection is based on the design of cost function. Nevertheless, how prosody can effectively help to select units remains a problem.

The use of prosody in a unit selection system is highly desirable. Some previous work usually used symbolic prosody, which is discrete description of prosody. The symbolic representation of prosody cannot give a fine distinction of prosody of units. Therefore, the best unit may be not selected in the unit selection process. Some other research work used parametric prosody. However, the parameters are not well defined and well normalized. In this work, we will incorporate parametric prosody into the unit selection process.

There were a few attempts in Chinese unit selection-based TTS. However, previous work for Chinese unit synthesis use simple break or template based prosody models. These considerations can improve speech a little in prosody. However, this improvement is sometimes only by chance. The lack of full prosody representation prevents it from generating speech of high quality. At least the following speech problems cannot be solved in previous approaches:

- Inappropriate duration: The duration of a speech unit is determined by the context where the unit appears. A TTS system without good prosody consideration may generate too long or too short units.

- Inappropriate loudness: Due to the same reason, some of the units may have a too loud or too soft sound compared with their neighboring units.

- Inappropriate pitch level: Sometimes, we can perceive some high pitch or low pitch sound in some TTS systems. This is mainly caused by incorrect pitch level.

- Unclear or wrong tone: There is no careful consideration of pitch contour of unit in speech. A unit with a correct tone in the original speech may change to a wrong tone when connected with other units from other context.

- Incorrect break: When a unit initially from the start position of a prosodic word is placed at a position of end of prosodic word (or vice versa), we can perceive an obvious unnaturalness. This is mainly caused by improper realization of break (or boundary effect).

In this work, we will integrate parametric prosody representation defined in previous chapter into the unit selection process (refer to 6.3.2). The aim of the work is to overcome the problems that occur in previous TTS systems.

6.2 Unit Selection Model in this Work

In this research, we use a unit selection-based model for speech synthesis. Different from various previous researches in Chinese and other languages, we integrate parametric prosody information into cost function and unit selection process. In addition, the cost functions are designed to suit the nature of Chinese language.

6.2.1 Unit Specifications

In this work, we choose syllable as our synthesis unit. The reason to choose syllable is that syllable is a relatively stable units. The coarticulation between syllables is relatively loose, while the coarticulation between sub-syllable units is very tight.

Each unit is specified by a feature vector, which will be used for matching in a unit selection process. Both the target units and units in inventory are described using this feature vector. The features describe the phonetic identity, phonetic context, break types around the unit, and prosody parameters of each unit. The features defined in this work includes the following:

- Phonetic identity of the unit: Using the pronunciation of the unit is to ensure that the candidate unit will have the same sound as the expected one. The pronunciation includes the initial, final and tone. There are 22 initials, 38 finals, and 5 tones defined in this work.

- Phonetic context: The coarticulation between two units is determined by the phonetic identity of its neighbors. The context of the unit will help to find the unit with similar context of a unit. The phonetic context consists of the initials, finals, and tones of previous and next units.

- Breaks around the unit: The break types before and after the unit. The prosodic properties of a unit before a break and after a break are quite different. The break type information is an important index to evaluate the similarity of two units. We defined five types of break, which are syllable break, word break, prosodic word break, minor phrase break, and major phrase break.

- Prosody parameters: The prosody parameters are a collection of parameters that describe the duration, pitch contour and energy of a unit.

The details of all the features are listed in Appendix B.

6.2.2 Corpus Coverage

For corpus-based speech synthesis, a large speech corpus should be built. The speech corpus consists of a large collection of utterances. The unit for the synthesis will be extracted from the corpus. It is ideal to cover context dependent units and prosody variants as much as possible. However, meeting the criteria needs very large speech corpus or sometimes is even impossible. As the cost of constructing a large corpus with high quality is very expensive, balance is usually made between coverage and size.

In this research, we built a corpus of around 38000 syllables. The corpus is designed to cover the frequently used context independent syllable and context dependent syllable as much as possible. As calculated in Chapter 3, the built corpus covers 99.8% of syllable occurrences in PKU People's Daily text corpus. When

context of unit is grouped by initial and final class, the speech corpus covers 76.8% of the unit occurrences in PKU text corpus. When loose coarticulation is grouped together, the speech corpus covers 90.4% of the unit occurrences in PKU text corpus. (Refer to Section 3.3.2 for details).

6.2.3 Implementation of Prosody by Unit Selection

The prosody is implemented in unit selection by selecting units with proper prosody properties. This is done by using prosody related subcosts in cost function. (Refer to 6.3.2) The selected units will be concatenated together to form a speech utterance. The speech of connected units itself exhibits prosody. No silence is inserted into speech to create a prosodic break in utterance. Tone is implemented by selecting units with proper pitch contour. Break is implemented by selecting proper boundary units.

6.2.4 Costs for Unit Selection

Cost function describes to what degree that the selected units deviate from perfect ones. The cost function mainly consists of unit cost and connection cost. Unit cost mainly concerns quality of the unit, while connection concerns the coarticulation effects between the two selected units.

(1) Unit Cost (C_{Unit})

Unit cost expresses the distance between the unit to select and the unit that we expect. In the selection of units, we first look for the units with the same syllable identity (initial, final and tone) as the expected units. As we expect to find the syllable that has same context situation as our target speech, the cost is to measure its distance from the perfect one. Unit cost is calculated by comparing the corresponding features of a unit or a sequence of units, as illustrated in Figure 6.2. In the figure, T_i is the target unit, U_i is the unit to be selected.

Here we classify unit subcosts into two categories, which are phonetic cost and prosodic cost. The subcosts define the phonetic and prosodic fitness of the units, which will be discussed later.

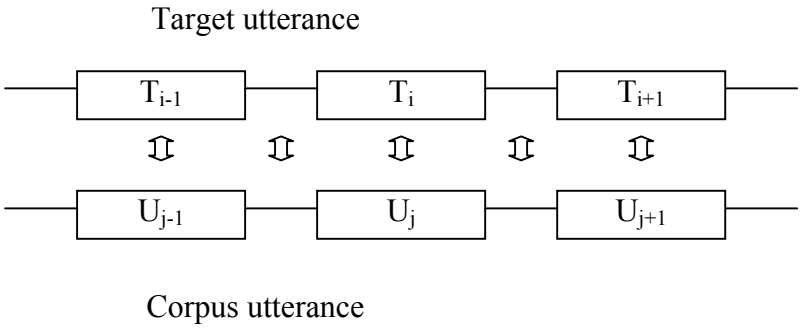


Figure 6.2 Illustration of unit cost calculation

(2) Connection Cost (C_{Conn})

When two selected best units from separate places are connected together, they do not necessarily match each other. Two successive units with sub-optimal unit cost may be preferable over two non-adjacent units with optimal unit cost.

The connection cost consists of two measures: coarticulatory continuity measure and prosodic continuity measure (Yi 1997). The First is inspired by the fact that certain phones spoken in succession exhibit a significant amount of coarticulation. Phone pairs with more perfect continuity in formants are more preferable to connect. Prosodic continuity compares the prosodic information of two connected syllables.

Figure 6.3 Direct calculation of connection cost

When two syllables are to be connected, if they were not spoken in succession, a connection cost must occur. The connection cost measures how much degrading in

the connection is caused when the pair of speech units comes from non-contiguous syllable constituents. The cost function can be calculated in two ways:

1. Directly calculated by calculating the spectrum continuity or prosody continuity between two units to be connected (as in Figure 6.3, in which units U_i and V_j are to be connected). This usually involves calculation of mismatch of acoustic or prosodic parameters.

2. Indirectly calculated by comparing the connected unit with its original neighbor in speech (as in Figure 6.4, in which units U_i and V_j are to be connected). This can be done by considering phonetic information. This work uses this way to describe connection cost.

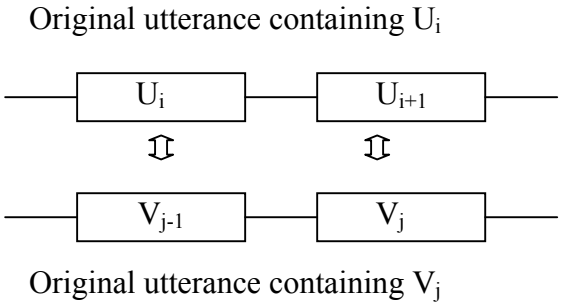


Figure 6.4 Indirect calculation of connection cost

Because some of the connections are more important (tight coarticulation or prosodically coherence) than the others are, we defined an importance factor for connection (which will be discussed later in 6.3.4).

6.2.5 Dynamic Programming

For each unit of the target speech, there are multiple speech units. The candidate units of all target units form a lattice. To find the path that has the lowest cost, a dynamic programming approach is needed. In this research, Viterbi algorithm is used to find the best path. The Viterbi search progress works in the following steps:

1. *Initialize C(0,1) = 0;*

2. *For i =1 to $N_{SeqUnit}$ do*

   a. *For j = 1 to N_{Cand}*

      Calculate unit cost C_{Unit} (j)

   b. *Sort units in ascending order of C_{Unit} (j), and keep the best M ones.*

   c. *For j =1 to N_{Path} do*

      For k =1 to M do

      $$C(i, jM+ k)= C(i-1, j) + W_{Unit}\ C_{Unit}(V_k) + W_{Con}C_{Con}(U_{i-1,j}, V_k)$$

   d. *Sort the paths in ascending order of C(i,1: jM+ k), keep the best N ones.*

3. *Back trace to find the best sequence that has a minimal cost value.*

4. *Output the sequence of units.*

where the meanings of the notations are as following:

$N_{SeqUnit}$: number of units in the sequence;

N_{Cand}: number of candidate units in current step;

N_{Path}: number of paths in previous step;

M: number of candidate units for further calculation in current step;

N : number of paths to keep in this step;

C(i,j) : accumulative cost of the *jth* path in the *ith* step;

V_k : the *kth* candidate in current step;

$U_{i,j}$: the *jth* selected unit in the ith step;

$C_{Unit}(V)$: the unit cost of unit *V;*

$C_{Con}(U, V)$: the connection cost between *U and V;*

W_{Unit}: weight for unit cost;

W_{con}: weight for connection cost.

6.3 Definition of the Cost Function

In this part, we will give the definition details of each subcost. In this work, the value of each single subcost is defined in a range from 0 to 100.

6.3.1 Phonetic Cost of Unit ($C_{Phonetic}$)

Phonetic context consists of final of previous syllable, initial of next syllable (or final of next syllable if the initial is null), tone of previous syllable, and tone of next syllable. The previous syllable and next syllable are considered due to the coarticulation effect and the interaction between them.

(1) Tone of Surrounding Syllables ($C_{ToneContext}$)

To calculate the cost, we calculate the cost for tone of previous syllable $C_{PrevTone}$ and cost for the tone of next syllable $C_{NextTone}$ respectively.

$$C_{PrevTone} = \begin{cases} 0, & \text{if } T_s = T_t \\ 50, & \text{if } T_s \neq T_t \end{cases} \qquad (6.1)$$

where T_t is the tone of the previous syllable of target syllable, and T_s is the tone of the previous syllable of a candidate syllable from inventory.

$$C_{NextTone} = \begin{cases} 0, & \text{if } T_s = T_t \\ 50, & \text{if } T_s \neq T_t \end{cases} \qquad (6.2)$$

where T_t is the tone of the next syllable of target syllable, and T_s is the tone of the next syllable of a candidate syllable from inventory.

Therefore the total subcost is

$$C_{ToneContext} = C_{PrevTone} + C_{NextTone} \qquad (6.3)$$

(2) Pronunciation of Context Syllables ($C_{PronContext}$)

To calculate this cost, we calculate the cost for the previous syllable $C_{PrevSyllable}$ and cost for the next syllable $C_{NextSyllable}$ respectively.

The cost $C_{PrevSyllable}$ is defined as:

$$C_{\mathrm{Pr}evSyllable} = \begin{cases} 0, & \text{if } F_s = F_t \\ 10, & \text{if } F_s \neq F_t \text{ but } FC_s = FC_t \\ 50, & \text{if } F_s \neq F_t \text{ but } FC_s \neq FC_t \end{cases} \quad (6.4)$$

where F_t is the final ID of the previous syllable of the target syllable, F_s is the final ID of the previous syllable of a candidate syllable from inventory, FC_t is the final class ID of the previous syllable of target syllable, FC_s is the final class ID of the previous syllable of the candidate syllable from inventory. The final class is as defined in Section 3.2.3.

Note that the IDs are numbers that represent categories.

The cost $C_{NextSyllable}$ is defined as:

$$C_{NextSyllable} = \begin{cases} 0, & \text{if } \mathrm{L}_s = L_t \\ 10, & \text{if } L_s \neq L_t \text{ but } LC_s = LC_t \\ 50, & \text{if } L_s \neq L_t \text{ but } LC_s \neq LC_t \end{cases} \quad (6.5)$$

where L_t is the left side (Left side is the initial of the syllable. When the initial is null, it is the final of the syllable) ID of the next syllable of the target syllable, L_s is the final ID of the next syllable of a candidate syllable from inventory, LC_t is the ID of left side class of the next syllable of target syllable, LC_s is ID of the left side class of the previous syllable of the candidate syllable from inventory. The left side class is as defined in Section 3.2.3.

Therefore the total subcost is

$$C_{\mathrm{Pr}onContext} = C_{\mathrm{Pr}evSyllable} + C_{NextSyllable} \quad (6.6)$$

6.3.2 Prosodic Cost of Unit ($C_{Prosodic}$)

Prosodic cost is calculated by calculating several subcosts firstly. In this work, we calculated subcosts for prosodic word breaks around the unit and prosody parameters of the unit.

Although the cost for prosodic break may be partly reflected in prosody parameters, we include it in the cost to give it more attention. We consider whether the unit is a prosodic word boundary or not because prosodic word is one of the most important factors for predicting prosody parameters.

(1) Break around the Syllable (C_{Break})

To calculate the cost, we calculate the cost for break before the syllable $C_{PrevBreak}$ and cost for the break after the syllable $C_{NextBreak}$ respectively.

$$C_{PrevBreak} = \begin{cases} 0, & \text{if } B_s = B_t \\ 50, & \text{if } B_s \neq B_t \end{cases} \tag{6.7}$$

where B_t is the break type (1: prosodic word break, 0: not a prosodic word break) before the target syllable, and B_s is the break before the candidate syllable from inventory.

$$C_{NextBreak} = \begin{cases} 0, & \text{if } B_s = B_t \\ 50, & \text{if } B_s \neq B_t \end{cases} \tag{6.8}$$

where B_t is the break type after the target syllable, and B_s is the break after the candidate syllable from inventory.

The total cost is

$$C_{Break} = C_{PrevBreak} + C_{NextBreak} \tag{6.9}$$

(2) Prosody Parameters ($C_{ProsodyParam}$)

Prosody parameters are predicted in Chapter 6. Here, we define a cost to account for all the prosody parameters. In this research, the parameters of a syllable defined for the cost calculation includes 10 parameters as determined in Section 5.6.1.

The calculation of the prosodic cost is defined as following. In the prediction of prosody parameters in last chapter, we obtain not only the values of prosody parameters we expect but also a value of standard deviation of the sample points falling into the corresponding leaf nodes of the regression tree.

The two values together give an accurate prediction of prosody parameters. The prosodic value gives the expected parameters, while the standard deviation reflects the accuracy of the value. Suppose the predicted prosody parameters are represented using vector T.

$$T = (t_1, t_2, ...t_{10}) \tag{6.10}$$

The corresponding standard deviations are presented using vector D.

$$D = (d_1, d_2, ...d_{10}) \tag{6.11}$$

The prosody parameters of a unit from inventory are represented using vector S.

$$S = (s_1, s_2, ...s_8) \tag{6.12}$$

The cost is calculated using

$$c = \sum_{i=1}^{10} (w_i \mid t_i - s_i \mid / d_i) \tag{6.13}$$

$$C_{\Pr osodyParam} = \begin{cases} 5c, & \text{if } c < 20 \\ 100, & \text{if } c \geq 20 \end{cases} \tag{6.14}$$

where w_i is the weight for each parameter.

6.3.3 Smoothness Cost between Two Units (C_{Smooth})

Suppose X, Y, P and Q are speech units as illustrated in Figure 6.5. X and Y are succeeding units in original speech, and P and Q are succeeding units in original speech as well. X and Q are to be connected in the synthetic speech as shown in Figure 6.5.

To calculate the connection cost between the two selected units that will be connected, we used the following features for each connection:

(1) Perfectly Connected (C_{Succ})

If the two selected syllables to be connected are originally succeeding units (X and P are the same unit) in the speech corpus, the cost should be zero. Otherwise (X and P are not the same), the cost is 100.

$$C_{succ} = \begin{cases} 0, & \text{if } X = P \\ 100, & \text{if } X \neq P \end{cases} \qquad (6.15)$$

(2) Tone Context ($C_{ToneConn}$)

To make the connected speech smooth, it is expected that the neighbors of the selected unit in the synthetic speech have same tones as those in the real corpus. The cost is calculated as:

$$C_{ToneConn} = \begin{cases} 0, & \text{if } T_x = T_P \text{and } T_Y = T_Q \\ 50, & \text{if } T_X \neq T_P \text{and } T_Y = T_Q \\ 50, & \text{if } T_X = T_P \text{and } T_Y \neq T_Q \\ 100, & \text{if } T_X \neq T_P \text{and } T_Y \neq T_Q \end{cases} \qquad (6.16)$$

where T_X, T_Y, T_P and T_Q are tones of units X, Y, P and Q.

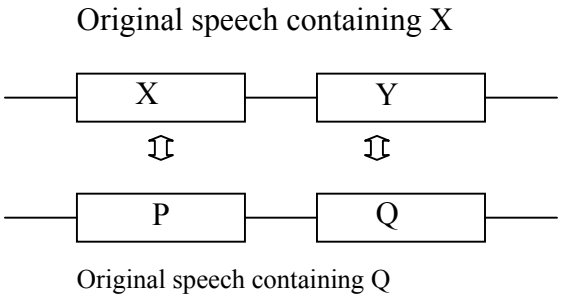


Figure 6.5 Connection cost calculation

(3) Phonetic Context ($C_{EdgeConn}$)

Each syllable has a left edge and a right edge (refer to Section 3.2.3). The left edge is classified by initial (if it is a null initial, the final is used). Right edge is identified by final.

$$C_{LeftSyl} = \begin{cases} 0, & \text{if } F_X = F_P \\ 10, & \text{if } F_X \neq F_P \text{ and } FC_X = FC_P \\ 50, & \text{if } F_X \neq F_P \text{ and } FC_X \neq FC_P \end{cases} \qquad (6.17)$$

where F_X is the final of the previous syllable of unit X, F_P is the final of unit P, FC_X is the final class of unit X, FC_P is the final class of the unit P.

$$C_{RightSyl} = \begin{cases} 0, & \text{if } L_Y = L_Q \\ 10, & \text{if } L_Y \neq L_Q \text{ and } LC_Y = LC_Q \\ 50, & \text{if } L_Y \neq L_Q \text{ but } LC_Y \neq LC_Q \end{cases} \qquad (6.18)$$

where L_Y is the left side class of unit Y, L_Q is the left side class of unit Q, LC_Y is the left side class of unit Y, LC_Q is the left side class of the unit Q.

The total subcost is calculated as:

$$C_{EdgeConn} = C_{LeftSyl} + C_{RightSyl} \qquad (6.19)$$

6.3.4 Connection Importance Factor Between Two Units (I_{Conn}).

As we are considering the connection between two units, first we need to have a look at the types of connection between the two syllables. There are different coarticulation degrees for different connection types. This considers two factors:

- Break types between syllables
- Coarticulation types between two syllables.

In this part, break type takes two values, which are existence or absence of a break. A break exists after a syllable when there is a prosodic word break, phrase break or major phrase break.

We define three types of coarticulation in Section 3.2.4. When two units are succeeding units in an utterance, the coarticulation is determined by pronunciation of the second unit (Wu et al. 2001). For different connection types, the connection cost should be given to different weights. Those tight connections should be strengthened and loose connection should be given more flexibility to select units that are not smoothly connected. This connection importance is a weight factor in the whole cost. The connection importance factor is defined as:

$$I_{Conn} = \begin{cases} 0.1, & \text{if } B(U,V) = 1 \\ 0.3, & \text{if } B(U,V) = 0 \text{ and } T(U,V) = 0 \\ 0.7, & \text{if } B(U,V) = 0 \text{ and } T(U,V) = 1 \\ 1.0, & \text{if } B(U,V) = 0 \text{ and } T(U,V) = 2 \end{cases} \tag{6.20}$$

where B (U,V) is the break type between U and V, the value 0 and 1 mean there is a prosodic word break or not between U and V respectively, T (U,V) is the coarticulation degree between U and V, 0, 1 and 2 means loose, intermediate and tight coarticulation respectively.

6.3.5 Total Cost

Total unit cost is calculated as:

$$C_{Phonetic} = W_{ToneContext} C_{ToneContext} + W_{\Pr onContext} C_{\Pr onContext} \quad (6.21)$$

$$C_{\Pr osodic} = W_{Break} C_{Break} + W_{\Pr osodyParam} C_{\Pr osodyParam} \quad (6.22)$$

$$C_{Unit} = C_{Phonetic} + C_{\Pr osodic} \quad (6.23)$$

where $W_{ToneContext}$, $W_{PronContext}$, W_{Break},, *and* $W_{ProsodyParam}$ are weights for the corresponding subcosts respectively.

Total connection cost is calculated as:

$$C_{Smooth} = W_{SuccUnit} C_{SuccUnit} + W_{CToneConn} C_{CToneConn} + W_{CEdgeConn} C_{CEdgeConn} \quad (6.24)$$

$$C_{Connection} = C_{Smooth} I_{Conn} \quad (6.25)$$

where $W_{SuccUnit,}$ $W_{ToneConn}$ and $W_{EdgeConn}$ are weights for the corresponding subcosts respectively.

Suppose a sequence of *n* units is selected for a target sequence of *n* units. The total cost is calculated with the following function.

$$C_{Total} = \sum_{i=1}^{n} C_{Unit}(i) + \sum_{i=0}^{n} C_{Connection}(i) \quad (6.26)$$

where the C_{Total} is total cost for the selected unit sequence, $C_{Unit}(i)$ is the unit cost of unit *i*, $C_{Connection}(i)$ is the connection cost between unit *i* and unit *i+1*. Unit 0 and n+1 are defined as start and end symbol to indicate start and end of utterance.

6.3.6 Weight Determination

The total cost of a sequence of units is a weighted sum of the unit cost and connection cost. The unit cost and connection cost are both weighted sum of sub-costs. Determining the weights is important for the general performance of the whole system. Unfortunately, it is hard to find an objective way to compare the quality of speech

utterances generated by using different weight settings. Therefore, we need to have some alternatives to determine the weights.

In this research, the weights are mainly determined by human based on knowledge and informal perception test. First, a set of weight values is assigned to each weight. Then the weights are adjusted to make the generated speech better.

(1) Initial Weights

The initial weights of unit cost are given according to the importance of the factor based on our knowledge. For the determination of costs, we follow the following rules:

- Cost of phonetic context ($W_{PronContext,}$) has larger weight than that of tone context ($W_{ToneContext,}$), boundary syllable (W_{Break}). The reason is that the phonetic context ensures the coarticulation of the syllable; while the tone context and boundary syllable type mainly determine prosody. The prosody is also contained prosody parameters.

- Cost of prosody parameters ($W_{ProsodyParam,}$) has a similar weight value to that of the phonetic context ($W_{PronContext,}$) because we want to give equal importance to them initially.

- The weight of cost of original connection ($W_{SuccUnit}$) is given a higher value than the others are. This favors selecting long speech segments.

(2) Weight Tuning

The tuning of weight is done by informal listening test. To make the adjustment of weights based on informal listening test more effective and meaningful, testing text is designed to evaluate the speech quality during adjustment of weights. The testing text consists of two parts:

1. Text has enough words that do not appear in the speech scripts. This is to test whether the generated speech has good prosody. The use of new words is to

ensure that the selected units have suitable prosody not because they happen to be selected from a unit of a same word.

2. Text consists of enough units, between which there are tight connections. This is to test whether the connection between units is well considered in the selected units.

The weights are adjusted to make the general speech quality is the best. Although this is not a formal testing, the weight can be adjusted to generate relatively good speech quality. Finally, the weights are adjusted as shown in Table 6.1.

Weight	Value
$W_{ToneContext,}$	0.5
$W_{PronContext,}$	2.0
$W_{Break,}$	1.0
$W_{ProsodyParam,}$	1.0
$W_{SuccUnit}$	0.5
$W_{ToneConn}$	0.3
$W_{EdgeConn}$	0.3

Table 6.1 Final weights in the cost function

Note that there are possibly different ways to set these weights. This setting is only one of them. This setting may not be the best one. However, it is enough to evaluate the performance of our prosody description.

6.4 Summary

In this chapter, we describe how prosody is adopted in the cost function. We describe the unit selection model and cost scheme used in this work. The general cost is divided into two main parts, which are unit cost and connection cost. The unit cost is further divided into phonetic cost and prosodic cost. We also define a connection weight for the connection cost. The procedure of weight tuning is also described.

The evaluation of the TTS system will be carried out in Chapter 7.

Chapter 7 Evaluation

In this chapter, we evaluate the performance of the TTS system when the prosody parameters are applied into the unit selection based synthesis. First, we highlight the key issues in the evaluation. Then the evaluation of the proposed system is carried out from different aspects.

7.1 Introduction of Speech Quality Evaluation

Evaluation of synthetic speech is difficult because the quality of speech should eventually be judged by human perception. Therefore, there is no directly automatic approach for testing like in speech recognition, in which recognition result can be compared with standard result automatically.

In the evaluation of speech quality, we are concerned about two important aspects, which are intelligibility and naturalness. Intelligibility means whether the speech is clear enough to convey the meanings that we intend to transmit. Naturalness, however, means whether the speech is pleasant to listeners.

The evaluation of synthetic speech is usually done by subjective listening test with a response set of syllables, words or sentences. Many approaches have been used in previous research for speech quality evaluations. We list some of the popular approaches in the following.

7.1.1 Segmental Unit Test

The testing material is usually focused on consonants, because they are easily confused. Many consonants are short or weak in speech. For example, in English, nasalized consonants (/m/, /n/, /ng/) are usually considered problematic. (Carlson et al. 1990). Some high frequency consonants (/f/, /th/, /s/) sometimes sound similar.

Diagnostic Rhyme Test (DRT), which was introduced by Fairbank in 1958, uses a set of isolated words to test consonant intelligibility in initial position (Goldstein 1995, Logan et al. 1989). The test consists of 96 word pairs, which differ by a single acoustic feature in the initial consonant. Word pairs are chosen to evaluate the six phonetic characteristics of speech.

Modified Rhyme Test (MRT), an extension to DRT, tests for both initial and final consonants apprehension (Logan et al. 1989, Goldstein 1995). The test consists of 50 sets of one-syllable words, which makes a total set of 300 words. In listening test, a word is played and listener is asked to make a multiple-choice answer for what he hears.

There are other variations of the method that use constructed syllable lists, word lists or nonsense word lists to evaluate different aspects of speech quality.

This kind of testing mainly concerns the intelligibility of speech segments. The designed testing set is language dependent.

7.1.2 Sentence Level Test

Several sets of sentences have been developed to evaluate the comprehension of the synthetic speech. Unlike in segmental tests, incorrectly perceived units can be corrected by context information in sentence.

Harvard Psychoacoustic Sentences is a closed set of 100 sentences developed to test the word intelligibility in sentence context (Allen et al. 1987). However, using a fixed set of sentences, learning effect is very problematic. Therefore, repeated experiments cannot be made. In addition, the words can be guessed from context.

Haskins Sentences uses texts in which missed items cannot be concluded from their contexts. However, a fixed sentence cannot be repeatedly used for test due to the learning effect.

Semantically unpredictable sentence test (SUS-test) is also a sentence level test (Goldstein 1995, Pols et al. 1992). The words to be tested are selected randomly from

a pre-defined list of possible candidates. The test contains five grammatical structures. In actual test, 50 sentences are generated and played in random order to test the subjects. This test is not sensitive to learning effect.

These tests are intended to test the intelligibility at sentence level. Note that the designed sentence sets are for English.

7.1.3 Overall Test

Mean opinion score (MOS) method is widely used to evaluate speech quality in speech transmission and speech synthesis (Goldstein 1995). MOS approach is to ask listeners to score each utterance. The average reflects the quality of speech.

This approach can be used to evaluate the general quality of speech or the quality of some specific features, for example, naturalness, intelligibility, prosody, etc. Sometimes, reference speech utterances are given as a guideline for scoring. However, due to the perceptual multidimensionality of speech (Sproat 1997), which means that there are usually different features in a speech utterance, listeners may focus their interests on different features for different utterances. The perceptual multidimensionality makes the use of reference speech ineffective. Therefore, in many MOS tests, a scale of five levels is given. However, the speakers are asked to score the speech utterances based on their own judgment. MOS test is usually used for relative listening test. That means, it is suitable for comparing two algorithms. It is usually meaningless to test one system alone using MOS test.

7.1.4 Objective Evaluation

All the above-mentioned approaches involve human listening of the speech utterances. Therefore, they are all subjective evaluation approaches. There were also attempts to use objective testing approaches.

Objective methods, such as Articulation Index (AI) or Speech Transmission Index (STI) are used to evaluate speech quality in speech transmission (Pols et al. 1992). These methods are unsuitable for synthetic speech because it is not possible to

establish a unique reference speech. However, some attempts are made to objectively evaluate the quality of speech in concatenative speech synthesis (Boefard et al.1993).

A typical idea for this method is to evaluate the speech quality by comparing the synthetic speech with a standard template. Some natural speech utterances from inventory are usually held up as the standard template speech. The comparison is done by comparing two sequences of speech features with dynamic time warping (DTW).

Although these approaches have been applied by some work, the main problems are: (1) The distance measures for comparing speech utterances do not necessarily reflect the perceptual differences of speeches. (2) The features used for evaluation do not contain enough prosody information. For example, duration information is ignored by using DTW. Pitch information is usually omitted in the features.

Therefore, objective testing approaches are usually useful for testing coarticulation effect but not for testing prosody effects. We have to rely on subjective listening test in this work.

7.2 Evaluation of Speech Quality

In this work, we evaluate the performance of the TTS from the following aspects:

- The performance of the parametric representation in synthesis
- The accuracy of the realization of prosodic effects (tone and break in this work)
- The quality of the generated speech (intelligibility and naturalness)

7.2.1 Testing Problem of this Work

The evaluation task for corpus based unit selection approach has some major differences from other synthesis approaches.

The previous testing approaches for segmental units are suitable for testing signal processing based approach, in which the same unit is usually generated from the same speech template unit. Therefore, if listening test shows that a unit is intelligible, the

same unit in a different occasion is usually intelligible as well. In such case, testing of a unit is in fact a complete test of the unit in different situations.

However, things are different in corpus-based unit selection approach because, in corpus-based synthesis, different occurrences of units often come from different source units. If one unit is intelligible, we cannot draw the conclusion that the same unit will be intelligible in different occasions.

Since the possible text of the language is an infinite set, we understand that any text for testing can only cover a very small part of this language. To better evaluate the quality of the speech, we need to design a text that has a good coverage of the language to some degree.

We also have some considerations of subjective listening test. Subjective testing is usually prone to error. To make the test more reliable, there should be enough observations. Therefore, it is expected that the testing units are small. For example, it is more accurate to use syllables as testing objects rather than to use sentence as testing objects. In the signal processing based system, it is usually difficult to identify which unit is not good because almost all the units are of similar quality. However, in a unit selection based system, we can identify which unit is bad.

7.2.2 Evaluation Methods in this Work

To evaluate the performance of the proposed prosody scheme in the unit selection process, we design some experiments to perform the tests.

(1) Evaluation of Cost Functions

In this work, we are to evaluate how prosody helps to select the proper units in synthesis process. Instead of judging the quality of a complete sentence, we judge the quality of each unit. This gives a more subtle comparison on two sets of speech samples. The unit level listening test is more objective than comparing two utterances. When conducting listening test, we ask listener to count the units that are considered not a good candidate of the expected unit. We define rate of inappropriate unit (RIU) to evaluate the synthetic speech. RIU is defined as the percentage of inappropriate

units among all units of the generated speech. The speech quality is better if this RIU value is smaller.

(2) Accuracy of Break and Tone

Break and tone are two of the most important perceptual prosody elements of Chinese speech. The information of break and tone are first derived from the input text. Then we convert all input information into prosody parameters. Finally, we implement all the effects by using unit selection approach. We are interested to know how well the break and the tone are preserved in the final speech after such transformations.

In the unit selection based synthesis approach in this work (Chapter 6), unit is defined as a tonal syllable. That means, when we want to select a unit, we will select a unit with the same pronunciation and same tone as the target unit. Ideally, the tones of all the selected units should be 100% correct in the synthetic speech. However, this is not true. In reality, some selected units are perceived as another tone. The reasons for this phenomenon are: (1) The corpus is not perfect. So, the tones of some syllables are not fully realized during reading. For example, some weak syllable is changed to a neutral tone or something between neutral tone and the original tone. (2) Tone contour depends on context tone. A tone is heard correct in one place might be heard incorrect when placed in another context. (3) It is possible that there are some errors in labeling. For example, start and end positions of a unit may not be accurate; a syllable may be labeled a wrong tone, etc.

The problem of break is similar. The final effect of break depends on the selected unit. For example, if we need a unit from the start position of a prosodic word, but a unit from the end position of a prosodic word is selected, an incorrect break may be perceived in the final speech.

The accuracy of tone and break are calculated by counting the number of units that are not perceived as correct break and tone respectively. Note that in the calculation of accuracy, we only distinguish prosodic word break or not a prosodic word break. Minor phrase break and major phrase break is not considered. The accuracy of break and accuracy of tone are separately calculated in this work.

(3) Speech Quality Evaluation

The speech quality testing involves the testing of intelligibility and naturalness.

The intelligibility test in this work is to listen to some nonsense sentences and then calculate the percentage of correctly heard units. This ensures that the listeners understand the syllable from its sound rather than from the whole context.

Naturalness test is done using MOS testing in this work. The MOS approach in the work is to ask the listener to score each sentence based on a 5-level scale of general naturalness of speech. The grading scale is shown in Table 7.1. This is a subject-oriented test. Quality of human speech is graded as 5. A speech utterance is marked as 5 (Excellent) if the listener thinks it is as good as the human speech. If a listener thinks a speech is good but is still not as good as human speech, it is marked as 4 (Good). If a speech is not so good but is acceptable, it is marked as 3 (Fair). If the listener thinks the speech is very bad, it is scored as 1 (Bad). Less bad ones are marked as 2 (Poor).

Naturalness	Excellent	Good	Fair	Poor	Bad
Score	5	4	3	2	1

Table 7.1 MOS scores for listening test.

(4) Reliability Consideration

Because subjective tests are involved in this evaluation, there is a problem of how to make the testing more reliable.

The main consideration is that how many listeners and how many listening material should be used in the tests. Using more listeners and more testing sentences improves reliability, but also increases the expenses. According to literature (Speechworks, 2002), for listening test, 10 subjects listening to 40 to 50 sentences from each system provides a good balance between cost and reliability of the result.

From statistics, adequate sample size depends on the confidence level required and the significance degree of the hypothesis to be tested. For a certain confidence

level, in the comparison of two populations, if the difference is small, the sample size should be large enough. Otherwise, a small sample size is usually enough.

7.2.3 Testing Material Selection

One of the issues in evaluating the synthetic speech is what testing text should be used in testing. Because natural language is an infinity set, any testing text is just infinitesimal. Therefore, we can only test a small part of real world text. To make test more reasonable, we have to design text to cover main problems in synthetic speech. In this work, text for testing naturalness mainly concerns the coverage of context dependent units. Testing text for intelligibility concerns the coverage of distinct syllables.

To select text for general speech quality testing, we use a greedy algorithm. The algorithm selects sentences from the PKU People's Daily corpus. The algorithm of selecting sentences can be described as the following algorithm.

1. Initialization
    i. Let S_0 be sentence set for selection;
    ii. Let T_0={};
2. Selection: for i = 1 to n do
    for all s in S_i do t= argmax F(s)
    $T_i = T_{i-1} + t$;
    $S_{i+1} = S_i - t$;
    F(t) = 0;
3. Output T_n

where S_i is the candidate sentence set for selection in i-th step, T_i is the selected sentence set in the i-th step, n is the number of sentence to be selected, and F(s) is the

sum of relative frequency in PKU corpus for all the units in sentence s. Note that F(t) is assigned zero if t has been selected in T_i.

The idea of the above algorithm is to select sentence one by one from a large set. Each selected sentence need to best cover the units that are not covered by previously selected sentences.

7.3 Experiments

In the following experiments, we will (1) select text for listening test; (2) evaluate the performance of our prosody representation; (3) evaluate the accuracy of the implementation of break and tone in synthesized speech; (4) evaluate the intelligibility of generated speech; (5) evaluate naturalness of the generated speech; (5) test the speed of the TTS system.

7.3.1 Testing Text Selection

Testing text of the listening test is selected from PKU People's Daily corpus (Yu et al, 2002). Frequencies of context dependent units are calculated as described in Section 3.2.4. Context dependent unit is defined by considering the following: (1)The left context are grouped by the final class of the previous syllable; (2) The right context are grouped by the initial class of the next syllable; (3) The units with loose connections to the previous syllable are grouped together; (4) The units with loose connections to the next syllable are grouped together.

First, we select the sentences with 8 to 12 characters as our candidate sentence set. Then, we use the algorithm described in 7.2.3 to select sentences. The selection result is as shown in Figure 7.1. In the figure, the x-axis shows the number of sentences selected, and the y-axis shows the percentage of covered units in PKU People's Daily Corpus. We found that when we select 1000 sentences, we can cover 94.6% of all the context dependent units in the People's Daily corpus. Finally, we select 100 sentences randomly from the first 1000 sentences as our testing sentence set. Note that we do not select the first 100 selected sentences, as we want to choose both frequent units and less frequent units for a fair testing. The selected sentences are as shown in

Appendix C. The 100 sentences consist of 1091 characters. The testing sentences will be used in some of the following experiments.

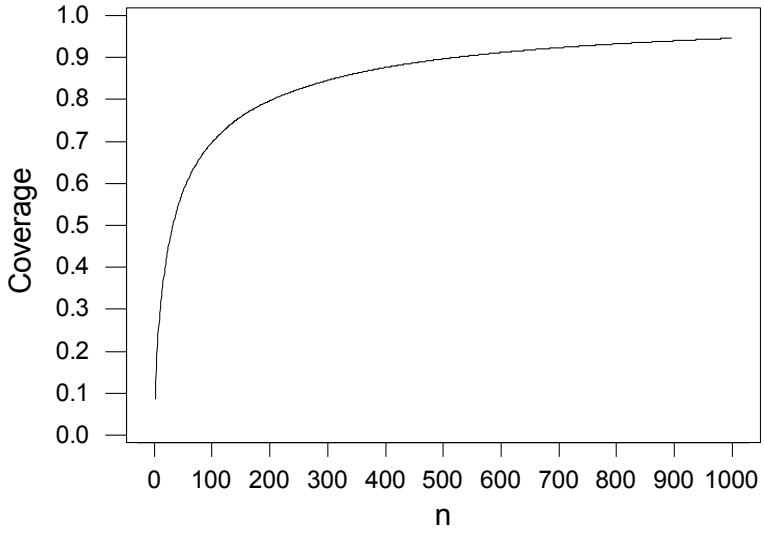


Figure 7.1 Text selection for listening test

7.3.2 Parametric Prosody vs Symbolic Prosody

In this experiment, we evaluate the performance of our prosodic representation. We compare the performance of the parametric prosody with that of symbolic representation, which is used by other research work.

We synthesize speech using three different ways. The difference between the approaches is in the calculation of cost function. The three methods are:

- **Method 1 (No prosody used):** In Method 1, the cost function only includes the phonetic cost and connection cost. No prosody and connection importance are included. That means, $C_{Prosodic} = 0$ and $I_{Conn} = 1$ in (7.24) and (7.26).

- **Method 2 (Symbolic prosody):** In Method 2, the cost function only includes the phonetic cost, prosodic cost, and connection cost. However, prosody is

only accounted for by using break types, i.e., $W_{ProsodyParam}$ =0. Also there is no connection importance factor included, i.e. I_{Conn} = 1 in (7.26).

- **Method 3 (Full prosody used):** It includes phonetic, prosodic connection cost, and connection importance factors, as described in Section 6.3.

A comparison of the subcosts that the cost functions use is shown in Table 7.2

	Cost function			
Methods	Phonetic	Smoothness	Prosody	Importance Factor
Method 1	Used	Used	Not used	Not used
Method 2	Used	Used	Break type	Not used
Method 3	Used	Used	Used	Used

Table 7.2 Methods used in cost test

Method 2 was adopted by previous work by Chu 2001, which is one of the representatives of state-of-art unit selection based TTS system. The difference is that their work used a larger corpus of around 200,000 units. However, there are only around 38,000 units in this work.

We synthesized the 100 sentences selected in 7.3.1 using the three methods. 20 native speakers of Chinese have participated in the listening test. They are asked to listening to the 3 sets of speech samples and count the units that are not considered good enough. The result is shown in Table 7.3. Experiment shows that, using method 1, the RIU (rate of inappropriate units) is 46.1%. Using methods, the RIU reduced to 32.2%. RIU is further reduced after using Method 3. This shows after using symbolic prosody, the naturalness is improved, and after using full parametric prosody representation, the naturalness is improved significantly.

Method	RIU Mean	StdDev
Method 1 (No Prosody)	46.1%	9.8%
Method 2 (Symbolic Prosody)	32.2%	6.7%
Method 3 (Parametric Prosody)	8.1%	4.2%

Table 7.3 Result of rate of inappropriate units(RIU)

In Method 1, no prosody is applied to the synthetic speech, but the smoothness between units is considered. Although some of the sentences are understandable, there are many prosodically inappropriate units found in the speech utterances. Main types of inappropriate units include:

1. Unclear tone. Some units sound like units with a different tone from its original tone. For example, a third tone may appear as a first tone, a first tone sounds like a second tone.

2. Unclear sound: Although each unit can be correctly identified in the original speech utterances in speech corpus, when they appear in a synthetic speech, the sound cannot be correctly recognized.

3. Incorrect break position: Some breaks are placed at wrong places. Some speech utterances seem incomplete.

4. Inappropriate duration: Some units sound too long or too short to be fitted in the speech utterances.

5. Inappropriate pitch level: Some units have a higher or lower pitch level than their neighboring units. We can hear a sharp rise or fall in pitch.

6. Inappropriate energy level: Some units appear louder or softer than their neighboring units. Volume change can be identified sometimes.

Some errors can be classified into more than one of the categories. All the above problems make it difficult to understand the synthetic speech or it makes listeners feel uncomfortable.

In method 2, when the break information is considered as a substitute of prosody, the number of inappropriate units is reduced. Most of type 3 errors are corrected, and some of the other type errors are reduced.

When the prosody parameters (method 3) are applied to the unit selection process, the number of the inappropriate units decreased significantly. After examining the inappropriate units, we found they are caused by the following reasons:

1. Wrong segmentation. This can be improved by improving word segmentation.

2. Wrong prosodic word break or phrase break. The number of errors of this kind can be reduced if the number of errors in POS tagging and prosodic word prediction can be reduced.

3. Incomplete variants in unit inventory. In some cases, no proper unit can be found. Improvements can be made if a larger inventory is used.

Among the three tested methods, Method 1 applies no prosody, however smoothness is considered. Method 2 is an implementation of cost defined by Chu et al (2001), which is one of the state-of-art Chinese TTS systems. Method 3 is our approach of cost design. The result shows that integrating parametric representation of prosody into the cost function greatly improves the quality of the synthetic speech.

7.3.3 Break and Tone Accuracy

This experiment is an extension of experiment in 7.3.2. We evaluate the accuracy of break and tone in synthetic speech. We used the synthetic speech in 7.3.2 as testing material.

(1) Break Experiment

In this experiment, we want to evaluate how well the breaks are implemented. We ask the 20 native listeners to listen to the synthetic speech, and count the breaks that are well implemented. The accuracy is recorded for comparison.

The result is as shown in Table 7.4. From the table, we can see that when no prosody is integrated, the accuracy of identifiable break is as low as 62.3%. When the symbolic representation is used, the accuracy rises to 87.2%. When the parametric prosody is applied, the accuracy of break is as high as 94.2%. This means that the use of the prosody parameters helps to improve the accuracy of the break placement.

We also note that, when no prosody is applied, the standard deviation is 10.3%. That means the number of correct breaks does not agree among listeners because the

breaks are not easy to be correctly identified. When the symbolic prosody is used, the standard deviation is 5.2%. When the parametric prosody is used, the standard deviation is 3.3%, which means that there is a more agreement on the identified breaks among listeners when prosody is used.

Note that the accuracy of break is 94.2%, which is higher than the accuracy of prosodic word break prediction in Chapter 4, where the highest accuracy is 91.65%. The accuracy of implemented break is higher than the predicted break. The reasons are: (1) In the break prediction, we compare the prediction result with the breaks labels in corpus. We have mentioned in Chapter 4 that the accuracy of the break prediction could be higher if we take into account the breaks that are different from corpus but are acceptable. (2) Some implemented breaks sound between a prosodic break and a none-break. So, they are accepted as correct breaks.

We also note that the accuracy of break from the symbolic prosody approach is 87.2%. This is lower than the accuracy of prosodic word prediction, which is 91.65%. The reason is that, in symbolic prosody approach, when an expected boundary unit cannot be found, a wrong boundary unit is used instead. This increases the break errors. The errors can be avoided in parametric prosody approach, in which, when there is no suitable boundary unit, a non-boundary unit with suitable prosody nature will be used. This explains why prosody parameter approach outperforms symbolic prosody approach in break implementation.

Method	Mean	StdDev
Without Prosody Parameter	62.3%	10.3%
With Symbolic Prosody	87.2%	5.2%
With Prosody Parameter	93.4%	3.3%

Table 7.4 Accuracy of break in speech

(2) Tone Experiment

In this experiment, we want to evaluate how well the tones are implemented. We ask the 20 native listeners to listen to the synthetic speech, and count the tones that are correctly implemented. The accuracy is recorded for comparison.

The result is as shown in Table 7.5. We note that when there is no prosody parameters used, the accuracy of tone is 78.3%. When the symbolic prosody is used, the accuracy is 86.1%. When the prosody parameters are applied, the accuracy rises to 97.1%. The standard deviation also falls from 5.2% to 4.5%, then to 1.3%. That means there is a high agreement of tone identity among listeners when the prosody parameters are applied.

The experiment shows that the use of the parametric prosody greatly helps to improve the tone accuracy.

Method	Mean	StdDev
Without prosody parameters	78.3%	5.2%
With Symbolic Prosody	86.1%	4.5%
With prosody parameters	97.1%	1.3%

Table 7.5 Result of correctly implemented tones

7.3.4 Quality of Synthetic Speech

The quality of speech is usually evaluated by two main indexes, which are intelligibility and naturalness.

We compare the performance of our system with that of others in the experiment. Two systems are selected for comparison. The first selected system is Microsoft SAPI 5.0, released in 2000. The reasons we choose SAPI for comparison are:

(1) SAPI is the most popular system available, and hence provides a good reference of synthetic speech quality. SAPI is a system with relatively good speech quality and good prosody.

(2) SAPI is a not a unit selection based speech synthesis system (it is based on LPC synthesis), in which speech can be generated with desired prosody precisely. However, our approach of unit selection can only select unit with fixed prosody properties. During unit selection, there might be a prosody mismatch between the selected units and expected unit. Therefore, this test can compare the different forms of implementation of prosody parameters.

Another system we want to compare with is Ifly TTS system. This system is developed by Ifly Company, which is a leading Chinese TTS provider in the world (http://www.iflytek.com). The system for comparison was released in 2002. The reasons to select this system are: (1) The system is generally considered among the best ones. (2) It represents the latest TTS technology. (3) It uses a unit selection based approach. This provides a similar ground for testing performance of my system.

We also compare the generated speech with that generated using my implementation of the approaches using symbolic prosody representation.

(1) Intelligibility

The intelligibility can be judged by the rate of recognized units (RRU). In this test, we selected 400 most frequently used syllables. Neutral tone (or tone 5) is not considered. There is an average of 55 variants of each syllable in the speech inventory. The 400 syllables cover 56.1% of all the syllables in the inventory. We first randomize the syllables and then construct 80 nonsense sentences, in which each sentence consists of 5 characters. The reason that we choose five characters is that it is difficult for listeners to remember too long meaningless syllable sequences.

When listening to the utterances generated by different approaches, there is a problem of learning effect. That means, listener may remember the content of the utterance. That will make the result of intelligibility test unreliable. To avoid this, we

generated 4 sets of sentences. The characters in each set have different orders. Each set of sentences is used for one approach. One set of generated testing text is as shown in Appendix D.

Synthesis approach	RRU Mean	Standard deviation
Symbolic Prosody	80.2%	6.4%
Microsoft SAPI 5.0 (2000)	83.4%	3.2%
Ifly (2002)	88.3%	3.8%
My System	91.2%	4.1%

Table 7.6 Result for intelligibility test (Rate of recognized units)

30 native speakers of Chinese participated in the listening test. Among them, there are 15 females and 15 males. Listeners are asked to listen to each utterance and record what they heard on paper. Then we compare the syllables they recorded with the original text and count the correctly recognized syllables.

The result of the testing is shown in Table 7.6. We can see that the intelligibility of my system is 91.2%. Using symbolic prosody achieves 80.2%. The intelligibility of SAPI is 83.4%. The intelligibility of Ifly is 88.3%. It shows that, in intelligibility, my system is better than Microsoft SAPI and the approach of using symbolic prosody. The intelligibility of my system is slightly better than that of Ifly.

Although the prosody parameters are designed to improve the naturalness, the experiment shows that they also help to improve intelligibility. The reason is that prosody has some relation with intelligibility. When units is improperly read or labeled in corpus construction process, it affects both the intelligibility and the prosody of the unit. Therefore, intelligibility deficiency can be also reflected by prosody.

(2) Naturalness

The naturalness is usually judged by MOS test. In this test, we use the 100 sentences selected in Section 7.3.1 as the testing set.

Synthesis approach	MOS Mean	Standard deviation of MOS
Symbolic Prosody	3.12	0.38
Microsoft SAPI 5.0 (2000)	3.41	0.41
Ifly (2002)	3.94	0.10
My System	4.21	0.23

Table 7.7 Result for naturalness test

30 native speakers of Chinese participated in the listening test. Among them, there are 15 females and 15 males. The listeners are asked to compare utterances generated by 4 approaches and score them. Because different voices are involved in evaluating these systems, we ask listeners to concentrate on prosody properties of the speech. The 4 utterances of each sentence is played one by one. However, to avoid listener developing a bias during listening, the order of synthesis approaches is randomized.

The MOS testing result is shown in Table 7.7. In the table, we see that my system has a MOS score of 4.21, which is higher than other approaches. We see that SAPI has better score (3.41) than symbolic prosody approach (3.12). This shows, although the voice of SAPI is not as good as symbolic prosody approach, the prosody of SAPI is better. The result also shows my system is better than Ifly system.

7.3.5 Speed of TTS system

The speed of a TTS system can be measured by the number of the syllables synthesized in one second or the time length of the speech generated in one second. The speed is tested on a PC with CPU of Intel Pentium-III 1000MHz and memory of 256M, the operating system is Window 2000 professional. We selected 200 sentences, which consists of 2312 syllables in the test.

(1) Speed for different beam widths

As the speed of the unit selection is largely dependent on the beam width (Number of best paths kept in each step, the variable N in the algorithm in 6.2.4) of Viterbi search.

In this test, we show the speed of the unit selection process for different beam width values.

N	S	R
1	151.3	38.9
2	93.1	23.9
3	63.8	16.4
4	47.3	12.1
5	39.9	10.3
6	36.1	9.3
7	31.1	8.0
8	28.1	7.2
9	25.6	6.6
10	22.6	5.8

Table 7.8 Speed of unit selection dependent on beam width

The result is shown in Table 7.8 and in Figure 7.2. In the table, N is the beam width of Viterbi search process. S is the number of syllables synthesized in one second. R is the speech length (measured in second) generated in one second.

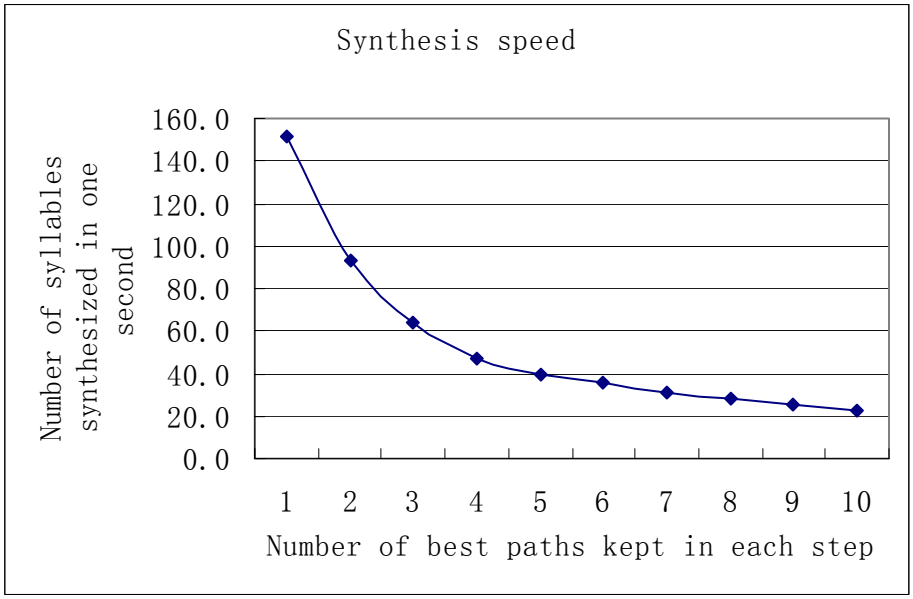


Figure 7.2 Speed of unit selection

In the table and in the figure, we see that the speed of unit selection drops with the increase of beam width. When the beam width is 10, the synthesis speed is 22.6 syllables or 5.8 seconds of speech per second. The speed is enough to be used in real

time TTS application. A larger value of N will allow a larger searching space. However, considering the speed of the system, we choose N=10 as our beam width. It should be mentioned that all the previous experiments were based on N=10.

System	Syllables/Second
My system	22.6
SAPI	154.1
Ifly	104.2

Table 7.9 Synthesis speed comparison

To understand the speed of the TTS system, we also synthesized the same testing text using Microsoft SAPI 5.0 and Ifly system. The speeds are compared in Table 7.9. The result shows that the SAPI has a speed of 154.1 syllables per second. We find that when the beam width is 1, the speed of our system is 151.3 syllables per second, which is compatible with SAPI. However, the speech quality for N=1 will not be as good as that when N =10. When N=10, SAPI has a speed of around 7.5 times of the speed of my system. The speed of Ifly is 104.2 syllables per seconds, which is 4.6 times of the speed of my system.

(2) Time breakdown of TTS System

There are three main parts in the TTS system, which are text analysis, prosody generation, and unit selection. The amounts of time used in text analysis, prosody generation, and unit selection are as shown in Table 7.10 and Figure 7.3.

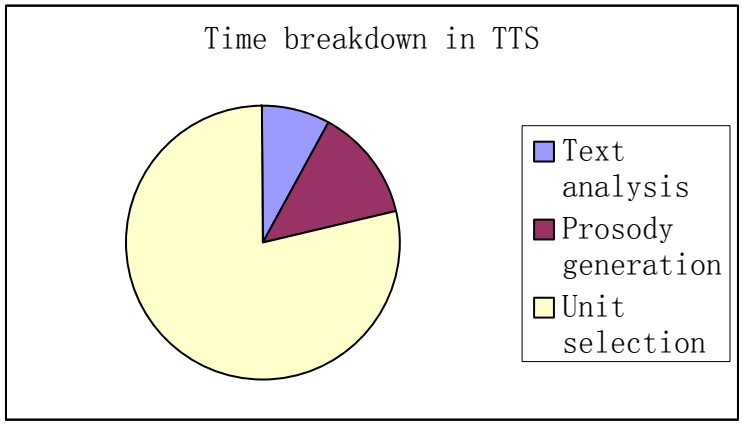


Figure 7.3 Time breakdown of the TTS

In the figure and in the table, we see that unit selection part takes most of the time in the whole TTS process. Therefore, improving the speed of unit selection will increase the TTS speed.

Component of TTS	Time percentage
Text analysis	8.1%
Prosody generation	13.2%
Unit selection	78.7%

Table 7.10 Time breakdown for TTS

Although my system can work for real-time use, we should note that my system is an experimental system. In my system, many of the data are stored in files instead of staying in the memory; the algorithms are not optimized; the data are not indexed. Therefore, there is space for improvement, especially for the unit selection part.

7.4 Discussion

From above experiments, we find the following:

1. Applying prosody parameters in unit selection-based synthesis can improve speech quality significantly.

2. The perceptual prosody elements, tone and break, are well implemented in the final speech.

3. The intelligibility and naturalness of the synthesized speech using the prosody parameters are much higher than that is generated by symbolic prosody, or SAPI.

4. The intelligibility is comparable with Ifly TTS system. The naturalness of speech generated by my system is higher than that by Ifly system.

5. The TTS system can be used for real-time use. Most of the time consumption is at the unit selection part.

The experiments show that the unit selection approach with integration of prosody generates speech with very high quality.

Although the TTS system can generate good speech, we should mention some disadvantages. The main disadvantages include:

- One of the advantages of application of the prosodic parameters is that boundary affects are well implemented in final speech. However, this is also sometimes a disadvantage in a real TTS system because the synthetic speech is sensitive to wrong break placements. When there are errors in break prediction, the wrong breaks are also truthfully implemented in the final speech. The errors in break prediction can be easily perceived. However, this problem can be alleviated by improving the models for break prediction.

- Unit selection-based approach needs a large corpus to work. Although the general speech quality is high, there are chances that when there is no proper variant of a needed unit. In such a case, the quality of some part of the speech may be bad. This makes the system unstable in some rare cases. The traditional signal processing approach, on the other hand, generates stable quality of speech, although the speech is machine-like.

- In a real system, to cover more variants of units, the corpus has to be very large. The recording, labeling and manual verification work in building such a system makes this approach very expensive. In a working system, it needs large storage to hold the speech data. This makes the system too huge to work on computers with small memories.

- Unlike signal processing based approach, it is not easy to adjust the pitch level, speaking rate of the synthetic speech in a pure unit selection based system. Such modifications are however very easy in signal processing based synthesis (such as Microsoft SAPI).

7.5 Summary

In this chapter, we introduced the problems and approaches for evaluating synthetic speech. We designed evaluation approaches and a testing text. We evaluated the performance of the prosody parameters and the TTS system.

In the evaluation of the speech quality, I developed an approach to select a testing text, which better covers the language in testing. I designed a syllable level speech listening test approach, which provides better distinction ability than sentence level testings.

The experiments show that the use of the proposed parametric representation of prosody in unit selection based synthesis greatly improves the speech quality than using symbolic prosody information. The intelligibility and naturalness of the generated speech are much better than SAPI and the approach using symbolic prosody. The system can work in real-time applications.

Chapter 8 Conclusion

The final chapter summarizes the research in the thesis, lists the contribution of the author, and gives directions for future work.

8.1 Summary of the Research

This research is an investigation of the problem of prosody generation for Mandarin Chinese text-to-speech system. I mainly work on two issues of prosody: (1) The prediction of prosodic phrase breaks, especially the prediction of prosodic word break. (2) The design, evaluation, and selection of prosodic parameters for unit selection based synthesis approach.

This work uses a speech corpus read by a female professional speaker. During the evaluation of speech corpus, the problem of speech unit distribution of Chinese language is first investigated. The speech corpus is then evaluated to find that it is suitable for this work.

The problem of prosodic break is investigated. The factors that affect the performance of prosodic break are examined. Dependency models for break prediction are developed. The experiments show that the models produce better result than simple CART approaches.

The approaches of designing, evaluating, and selecting prosody parameters are given. Some prosody parameters are defined to suit the nature of Chinese speech and the approach of unit selection. The parameters defined in this work are intended to overcome the major speech problems in speech synthesis. We highlight the problems of correctly representing perceptual prosody information (break and tone) in this work. The defined parameters are examined from the views of statistics and recognition. A clustering approach is used to remove redundancy in the prosody parameter definition. The relationship between parameters and features for prediction is investigated.

In the unit selection-based synthesis, the defined parametric prosody expression is applied in the cost function. The cost function is designed to suit the needs of Chinese language. Some experiments are designed to better evaluate the system. The experiments show that the use of parametric prosody representation significantly improves the quality of speech.

8.2 Contributions

The major contribution of this work is on the prosody application in unit selection based synthesis. I developed an approach to design and apply parametric representation of prosody suitable for unit selection-based synthesis for Chinese. As far as I know, this is the first work that investigates the design of parametric representation of prosody in a unit selection-based synthesis (for Chinese or other languages). Using this approach, we can transmit information of perceptual effects (break and tone) from linguistic features to prosody parameters, and then implement these effects by unit selection. The intelligibility and the naturalness of speech are improved.

Although this work is done through building up a complete text-to-speech system, the contribution of this work is not limited to a Chinese TTS System. Specifically, main contributions are in the following aspects:

(1) Methodology

In this work, I proposed an approach to apply parametric prosody representation in a unit selection based synthesis process. This approach solved the following problems that encountered in unit selection based speech synthesis. (1) The approaches for evaluating prosodic parameters have been given. This helps to determine whether the parameters are sufficient to describe perceptual prosodic effects (e.g. tone and break). (2) The approach for determining final parameter set has been given. The approach can determine a parameter set, which is concise but sufficient. (3) Using a regression tree approach, the prosody model predict the prosodic parameter as well as the standard deviation of the class to which it belongs. This makes it possible to measure mismatch in unit selection based synthesis.

Generally, this work provides solutions for determining a set of prosodic parameters that are suitable for unit selection based synthesis. Meanwhile, the approach makes sure that the selected parameter set is sufficient but concise. The selected parameters describe not only the general prosody of speech but also the important perceptual prosodic effects. The proposed approach can be extended to other prosody properties of Chinese or other languages.

In the work of break prediction, I evaluated my models for the prediction of prosodic word break and minor phrase break. I found some ways to make improvements in predictions. The models can generate better prediction result than generally used CART approaches.

In the evaluation of speech corpus, I used some approaches to reduce the number of context dependent units. This solution reduces the number of context dependent units significantly. It makes building small speech inventory for text-to-speech synthesis possible.

In the evaluation of speech quality, I developed an approach to select testing set, which better covers the language in testing. I designed syllable level speech listening test approach, which provides better distinction ability than sentence level testing.

(2) Knowledge of the Chinese Speech

Because the work is done through building a Chinese TTS system, we achieved many findings during the building process. They are summarized as the following:

The statistical analysis shows that it is infeasible or even impossible to completely cover variants of unit in Chinese language. However, the problem of unit coverage can be alleviated by reducing the space of the units. We conclude that the corpus should usually be designed to balance between the corpus size and coverage of speech phenomena.

For the prosodic word prediction, we understand that the length of words and part of speech are important features for Chinese prosodic word break prediction. There is

a dependency between breaks, which helps to improve the accuracy of prosodic break prediction.

For the prosody description for Chinese, I discovered that energy contour (or its equivalent) help describe boundary units. I have discovered the relationship between prosody parameters and the features for prediction. This result helps understand the prosody parameters and features better. This is useful when building prosody models of different sizes, in which prosody model can be simplified by overlooking unimportant factors.

During the evaluation, a testing text is selected. It shows that it is possible to design a relative small testing set to test the speech of this language.

(3) Application

A complete text-to-speech system is obtained from this research. Therefore, the thesis can be used as a guide to build a practical text-to-speech system.

In the thesis, the approaches to predict prosodic word breaks and minor phrase breaks have been given. The features have been tested. The algorithms are also provided.

In the prosody parameter prediction, I defined a set of features for prediction. Through experiments, I determined a set of parameters that can be directly used in building prosody model for unit selection based speech synthesis approach.

In the unit selection based synthesis approach, the details of definition of cost function are provided. All these can be directly applied in a real Chinese TTS system.

8.3 Future Work

In the prosodic word prediction, wrong segmentation of words and wrong tagging part of speech may affect the accuracy of prediction result. Therefore, the problem of prosodic word may need to be considered with the problems of Chinese word segmentation. Some problems may be resolved at the stage of word segmentation.

In construction of a speech corpus, there are also some problems to be resolved. The labeling of speech corpus is a labor-intensive task. In this work, I used an automatic approach followed by manual checking. This manual labeling work is very slow. It is expected to have an approach to automatically make a good labeling without manual check.

In a labeled speech corpus, some units may not be good enough. For example, the sounds of some speech are not clear; some units cannot clearly cut out from its neighbors. How to eliminate these units from corpus needs more investigation. In this research, we have proposed approaches for recognizing tone and breaks. The recognition techniques could be used for inventory pruning.

The weight determination for the cost function in unit selection is important. However, there is no good method to resolve this problem now. The problem should be further investigated.

Bibliography

[1] Allen, J. Overview of Text-To-Speech Synthesis. In Furui, S. and Sondi, M. M., editors, Advances in Speech Signal Processing, Page 741-790. Marcel Dekker, 1991.

[2] Allen, J.; Hunnicutt, M. S. and Klatt, D. From Text to Speech: The MITalk System, Cambridge University Press, 1987.

[3] Anderson, M.; Pierrehumbert, J. and Liberman, M. Synthesis by Rule of English Intonation Patterns. In Proceedings of the ICASSP, pp 2.8.1-2.8.4, 1984.

[4] Bachenko, J. and Fitzpatrick, E. A Computational Grammar of Discourse-Neutral Prosodic Phrasing In English. Computational Linguistics, 16(3):155--170, 1990.

[5] Bagshaw, P.C. Unsupervised Training of Phone Duration and Energy Models for Text-To-Speech Synthesis. Pages 17-20 of: Proc. 5th. International Conference on Spoken Language Processing, vol. 2. Sydney, Australia. 1998.

[6] Beutnagel, M.; Conkie, A. and Syrdal A. Diphone Synthesis Using Unit Selection. In: The 3rd ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, NSW, Australia, Nov. 1998, Paper F.2 (R52).

[7] Beutnagel, Mark and Alistair Conkie. Interaction of Units in a Unit Selection Database. In Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary), Volume 3, Pages 1063-1066. 1999.

[8] Beutnagel, Mark; Mohri, Mehryar and Riley, Michael. Rapid Unit Selection from a Large Speech Corpus for Concatenative Speech Synthesis. In Eurospeech, Budapest, 1999.

[9] Black, A. and Campbell, N. Optimizing Selection of Units from Speech Databases for Concatenative Synthesis. In Proceedings of Eurospeech, pages 581--584, 1995.

[10] Black, A. and Hunt, A. Generating F0 Contours from ToBI Labels Using Linear Regression. In ICSLP96 volume 3, pp 1385-1388, Philadelphia, Penn., 1996.

[11] Black, A. and Lenzo, K. Optimal Data Selection for Unit Selection Synthesis. In 4th ESCA Workshop on Speech Synthesis, Scotland, 2001.

[12] Black, A. and Taylor, P. Automatically Clustering Similar Units for Unit Selection in Speech Synthesis. In Eurospeech97, volume 2, pages 601--604, Rhodes, Greece, 1997.

[13] Black, A. and Taylor, P. CHATR: A Generic Speech Synthesis System. In Proceedings of COLING94, Kyoto, Japan, 1994.

[14] Black, Alan and Taylor, Paul. Assigning Intonation Elements and Prosodic Phrasing For English Speech Synthesis from High Level Linguistic Input. In Proc. ICSLP `94, Yokohama, Japan, 1994b.

[15] Black, Alan; P. Taylor, and R. Caley. The festival speech synthesis system. http://www.cstr.ed.ac.uk/projects/festival.html, 1998.

[16] Breiman, L.; Friedman, J.; Olshen, R. and Stone, C. Classification and Regression Trees. Wadsworth and Brooks, Pacific Grove, CA., 1984.

[17] Buhmann, J. Vereechen, H. Fackrell, J. Martens, J. P. and Coile, B. Data driven intonation modeling of 6 languages, 2000.

[18] Bulyko, I., and Ostendorf, M., Joint Prosody Prediction and Unit Selection for Concatenative Speech Synthesis, In Proc. of ICASSP, 2001.

[19] Campbell, N. and Black, A. Prosody and the selection of source units for concatenative synthesis. In J. van Santen, R. Sproat, J. Olive, and J. Hirschberg, editors, Progress in speech synthesis, pages 279–282. Springer Verlag, 1996.

[20] Campbell, N. Processing a Speech Corpus for CHATR Synthesis, Proc. of ICSP'97, pp. 183--186, 1997.

[21] Campbell, Nick. Reducing the Size of a Speech Corpus for Concatenation Waveform Synthesis. Technical Publications, ATR Interpreting Telecommunications Research Laboratories, pages 90-91. 1999.

[22] Cao, Jianfen; Lv, Shinan; Yang, Yufang. Strategy and tactics on The Enhancement of Naturalness in Chinese TTS, ISCSLP, Beijing, 2000.

[23] Carlson R., Granstom B., Nord L. Evaluation and development of the KTH Text-to-Speech Systems on Segmental Level. Proceedings of ICASSP 90(1): 317-320, 1990.

[24] Chan, N. C. and Chan, C. Prosodic rules for connected Mandarin synthesis. J. Inform. Sci. Eng. 8, 261-281. 1992.

[25] Chao, Yuen Ren. A Grammar of Spoken Chinese. University of California Press, Berkeley, 1968.

[26] Chao, Yuen Ren. Tone and Intonation in Chinese. Bulletin of the Institute of History and Philology, Academia Sinica, Vol. 4, No. 2, pp. 121--134, 1933.

[27] Charpentier, F. and Moulines, E. Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis Using Diphones. In Proceedings EUROSPEECH'89, Paris, France, Volume 2, pp. 13--19. 1989.

[28] Chen, S. H. A Corpus-Based Prosodic Modeling Methods for Mandarin and Min-Nan Text-To-Speech Conversions. ISCSLP 2000, Beijing, 2000.

[29] Chen, S. H.; Hwang, S. H. and Wang, Y. R., An RNN-Based Prosodic Information Synthesizer For Mandarin Text-To-Speech. IEEE Trans. Speech Audio Processing. 6(3), 226-239. 1998.

[30] Chen, Weijun; Lin, Fuzong; Li, Jianmin and Zhang, Bo. A New Prosodic Phrasing Model for Chinese TTS Systems, NLPRS 2001, Taipei, 2001.

[31] Choi, John, Hsiao-Wuen Hon, Jean-Luc Lebrun, Sun-Pin Lee, Gareth Loudon, Viet-Hoang Phan, and Yogananthan S. Yanhui, A Software Based High Performance Mandarin Text-to-Speech System. In Proceedings of ROCLING VII, pp. 35--50, 1994.

[32] Chou, F. C. and Tseng, CV. Y. Corpus-Based Mandarin Text-To-Speech Synthesis with Contextual Syllabic Units Based on Phonetic Properties. In: Proc. ICASSP, pp. 893- 896, 1998.

[33] Chou, Fu-chiang, Tseng, Chiu-yu, and Lee, Lin-shan. Automatic Generation of Prosodic Structure for High Quality Mandarin Speech Synthesis. In Proceedings of the International Conference on Spoken Language Processing, (Philadelphia, USA), ICSLP, 1996.

[34] Chu, Min and Lv, Shinan. High Intelligibility and Naturalness Chinese TTS System and Prosodic Rules. In Proceedings of the XIII International Congress of Phonetic Sciences, (Stockholm), pp. 334--337, 1995.

[35] Chu, Min. Research on Chinese TTS System with High Intelligibility and Naturalness. Ph.D thesis, Institute of Acoustics, Academia Sinica, Beijing, China. 1995.

[36] Chu, Min; Peng, Hu and Chang, Eric. A Concatenative Mandarin TTS System without Prosody Model and Prosody Modification. Proc. of 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Perthshire, Scotland, August 29 - September 1, 2001.

[37] Chu, Min; Peng, Hu; Yang, Hongyun and Chang, Eric. Selecting Non-Uniform Units from a Very Large Corpus for Concatenative Speech Synthesizer. ICASSP2001, Salt Lake City, May 7-11, 2001.

[38] CISTC (Chinese IT Standardization Technical Committee), Chinese Internal Code Specification, Dec. 1995.

[39] Cover, T. and Thomas, J. *Elements of Information Theory*. John Wiley and Sons, Inc, 1991.

[40] Dixon, N. and Maxey, H. Terminal analog synthesis of continuous speech using the diphone method of segment assembly. IEEE transactions on Audio and Electro-acoustics, 16:40-50, 1968.

[41] Donovan, R. and Woodland, P. Improvements in an HMM-based speech synthesizer. In Eurospeech95, volume 1, pages 573–576, Madrid, Spain, 1995.

[42] Donovan, R. E. Trainable Speech Synthesis. PhD thesis, Cambridge Univ. Eng. Dept., June 1996.

[43] Donovan, R. The IBM trainable speech synthesis system, in ICSLP, December 1998, vol. 5, pp. 1703-1706.

[44] Dutoit, T. An Introduction to Text to Speech Synthesis. Kluwer Academic Publishers. 1997.

[45] Feng, Shengli. Interactions between Morphology Syntax and Prosody in Chinese. Peking University Press, Beijing, 1997.

[46] Flanagan, Jim. Speech analysis, synthesis and perception, Springer-Verlag., New York, 1972

[47] Fujio, S., Y. Sagisaka, and N. Higuchi. Prediction of Major Phrase Boundary Location and Pause Insertion Using a Stochastic Context-free Grammar, in Computing Prosody (Yoshinori Sagisaka, Nick Campbell, Norio Higuchi, editors), pp.271-284. 1996.

[48] Fujio, S., Y. Sagisaka, and N. Higuchi. Stochastic Modeling of Pause Insertion using Context-free Grammar. In Proceedings of the International Conference on Acoustic, Speech and Signal Processing, pp 604-607. 1995.

[49] Fujio, Shigery, Yoshinori Sagisaka and Norio Higuchi, Prediction of Prosodic Phrase Boundaries Using Stochastic Context-free Grammar, ICSLP, pp. 839-842, 1994.

[50] Fujisaki, H. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In Osamu Fujimura, editor, Vocal Fold physiology: Voice production, Mechanisms and functions, Raven, NY, pp: 347-355, 1988.

[51] Fujisaki, H. and Ohno, S. Analysis and Modeling of fundamental frequency contours of English utterances. In proceedings of Eurospeech, pp: 985-988, 1995.

[52] Fujisaki, H. and Sudo, H. A generative model for the prosody of connected speech in Japanese. Annual Report of Engineering Research Institute, 30:75--80, 1971.

[53] Fujisaki, H.; Hirose, K. and Lei, H., Prosody and Syntax in Spoken Sentences of Standard Chinese, Proc. ICSLP-92, 1, pp. 433--436, 1992.

[54] Gårding, Eva. Speech Act and Tonal Pattern in Standard Chinese: Constancy and variation. Phonetica 44, pp. 13-29. 1987.

[55] Goldstein M. Classification of Methods Used for Assessment of Text-to-Speech Systems According to the Demands Placed on the Listener. Speech Communication vol. 16: 225-244. 1995.

[56] He, Yang and Jin, Song. Intonations of Beijing dialect: An Experimental Exploration. Language Education and Research (in Chinese), 1992.1 pp 71-96. Beijing, China.

[57] Hirschberg, J. and Prieto, P. Training Intonation Phrase Rules Automatically for English and Spanish Text-to-speech. In Proc. ESCA Workshop on Speech Synthesis, pages 159--163, Mohonk, NY, 1994.

[58] Hon, H. W. et al. Towards large vocabulary Mandarin speech recognition. Proceedings of ICASSP 1994. pp:545-548.

[59] Hunt, J. and Black, A. Unit selection in a concatenative speech synthesis system using a large speech database. In ICASSP-96, volume 1, pages 373–376, Atlanta, Georgia, 1996.

[60] Hunt, J. Syntactic Influence on Prosodic Phrasing in the Framework of the Link Grammar. In Proc. European Conf. on Speech Communication and Technology, volume 2, pages 997-- 1000, Madrid, Spain, 1995.

[61] Hwang, Shaw-Hwa and Chen, Sin-Horng. A Prosodic Model of Mandarin Speech and its Application to Pitch Level Generation for Text-to-Speech. In Proceedings of IEEE ICASSP, Vol. 1, pp. 616-- 619, 1995.

[62] Hwang, Shaw-Hwa, Chen, Sin-Horng, Wang, Jih-Ru. A Mandarin Text-to-Speech System. International Journal of Computational Linguistics and Chinese Language Processing, Vol. 1, No. 1., pp. 87-100, 1996.

[63] Jilka, M. Mohler, G. and Dogil, H. Rules for the generation of ToBI-based American English intonation, Speech Communications, 28:83-108, 1999.

[64] Jin, Shunde, An Acoustic Study of Sentence Stress in Mandarin Chinese. PhD Thesis, Ohio State University, 1996.

[65] Klatt, D. H. Review of Text to Speech Conversion for English, Journal of the Acoustical Society of America, vol.82, no.3, pp.737-793, 1987.

[66] Kochanski, Gerg; Shih, Chilin and Jing, Hongyan. Hierarchical Structure and Word Strength Prediction of Mandarin Prosody, 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Scotland, 2001.

[67] Lee S. H. Tree-Based Modeling of Prosody for Korean TTS Systems. PhD thesis, Korea Advanced Institute of Science and Technology, 2000.

[68] Lee, J. C. Hang, D.G. Kim, S.H. and Sun, K.M. Energy contour generation for a Sentence using a neural network method. In proceedings of ICSLP 98. pp: 1991-1994, 1998.

[69] Lee, Lin-Shan; Tseng, Chiu-Yu and Hsieh, Ching-Jiang. Improved Tone Concatenation Rules in a Formant-based Chinese Text-to-Speech System. IEEE Transactions on Speech and Audio Processing, Vol. 1, No. 3, pp. 287--294, 1993.

[70] Lee, Lin-Shan; Tseng, Chiu-Yu and Ouh-young, Ming. The Synthesis Rules in a Chinese Text-to-Speech System. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 37, No. 9, pp. 1309--1320, 1989.

[71] Lee, Sangho and Oh, Yung-Hwan, Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS systems, Speech Communication, vol. 28, pp. 283-300, 1999.

[72] Li, Wei; Lin, Zhenhua; Hu, Yu; Wang, Renhua. A Statistical Method for Computing Candidate Unit Cost in Corpus Based Chinese Speech Synthesis System. In proceeding of International Conference on Chinese Computing, Singapore, 2001.

[73] Liao, R. R. Pitch contour formation in Mandarin Chinese: A study of tone and Intonation. PhD. dissertation, the Ohio State University, 1994.

[74] Liu, Qingfeng; Wang, Ren-hua; Ma, Zhongke and Yin, Bo. Design and Realization of a Chinese Speech Platform, Tianyin Huwang System. Communications of Chinese and Oriental Languages Information Processing Society 8 (2), pp. 211-220, 1998.

[75] Ljolje, A. and Fallside, F. Synthesis of natural sounding pitch contours in isolated utterances using hidden Markov models. IEEE transactions on acoustics, speech and signal processing. 34(5):1074-1079, 1986.

[76] Ljolje, Andrej; Hirschberg, Julia; and van Santen, Jan, P.H. Automatic Speech Segmentation for Concatenative Inventory Selection, Progress In Speech Synthesis, pages 304-311

[77] Logan J., Greene B., Pisoni D. Segmental Intelligibility of Synthetic Speech Produced by Rule. Journal of the Acoustical Society of America, JASA vol. 86 (2):566-581, 1989.

[78] Manning, C.D. and Schutze, H. Foundations of Statistical natural language processing. The MIT press, 1999.

[79] Mixdorff, H. and Fujisaki, H. Analysis of voice fundamental frequency contours of German utterances using a quantitative model. In proceedings of international conference on spoken language processing, pp:2231-2234, 1994.

[80] Möbius Bernd. Corpus-based speech synthesis: methods and challenges. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS 6 (4), 87-116, 2000.

[81] Monaghan, A.I.C. Phonological domains for Intonation in Speech Synthesis, Proceedings of Eurospeech 89, Paris, pp. 502-506, 1989.

[82] Moulines, E. and Charpentier, F. Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. Speech Communication 9,453-467, 1990.

[83] NLRM (National Language Reform Meeting or Quanguo Wenzi Gaige Huiyi). Resolution of meeting on problem of normalizing modern Chinese, in compilation of archives of meetings for modern Chinese normalization, Science Press, Beijing, 1955.

[84] O'Shaughnessy, D. Relationships between Syntax and Prosody for Speech Synthesis. Proceedings of the ESCA Tutorial Day on Speech Synthesis, Autrans (France), 39-42. 1990.

[85] Ostendorf, M. and Veilleux, N. A hierarchical stochastic model for automatic prediction of prosodic boundary location. Computational Linguistics, 20(1):27-54, 1994.

[86] Pierrehumbert, J. Synthesizing Intonation. Journal of the Acoustic Society of America, 70 (4), pp. 985-995, 1981.

[87] Pols L. SAM-partners. Multilingual Synthesis Evaluation Methods. Proceedings of ICSLP 92(1): 181-184. 1992.

[88] Qian, Yao; Chu, Min; Peng, Hu. Segmenting Unrestricted Chinese Text Into Prosodic Words Instead Of Lexical Words, ICASSP 2000.

[89] Riedi, M. P. Controlling segmental duration in Speech synthesis systems. PhD. Thesis, Swiss Federal Institute of Technology, 1998.

[90] Riley, M.D. Tree-based modeling of segmental duration. In G. Bailly, C. Benoit, and T. R. Sawallis, editors, Talking machines, Theories, models, designs. pp:287-304, Elseview Science, 1992.

[91] Ross K. and Ostendorf, M. A Dynamical system model for generating fundamental frequency for speech synthesis. IEEE Transaction on Speech and Audio Processing, 7(3):295-309, 1999.

[92] Ross, K. and Ostendorf, M. Prediction Of Abstract Prosodic Labels For Speech Synthesis. Computer Speech and Language, 10:155--185, 1996.

[93] Ross, K. N. Modeling intonation for speech synthesis. PhD thesis. Boston University. 1995.

[94] Sagisaka Y.; Kaiki N.; Iwahashi N. and Mimura. K. Unit Selection in A Concatenative Speech Synthesis System Using Large Speech Database. International Conference on Spoken Language Processing. Philadelphia, Oct, 1996.

[95] Sagisaka, Y.; Kaiki, N.; Iwahashi, N.; Mimura. K. ATR v-Talk speech synthesis system. International conference on Spoken Language Systems, Banff, Canada, 1992, pp. 483-486.

[96] Sagisaka, Yoshinori and Naoto Iwahashi. Objective optimization in algorithms for text-to-speech synthesis. In W. Bastiaan Kleijn and Kuldips K. Paliwal, editors, Speech Coding and Synthesis. Elsevier, Amsterdam, pp: 685-706, 1995.

[97] Sagisaka, Yoshinori. Speech synthesis by rule using an optimal selection of non-uniform synthesis units. In Proceedings of the IEEE ICASSP, New York, pp: 679-682, 1988.

[98] Santen, J. van. Combinatorial Issues in Text-To-Speech Synthesis. In Proceedings Eurospeech, Rhodos, Greece, 1997.

[99] Santen, J. van. Prosodic Modeling in Text-To-Speech Synthesis. In Proc. Eurospeech-97, 1997.

[100] Santen, Jan P. H. van. Assignment of Segmental Duration in Text-to-Speech Synthesis. Computer Speech and Language, Vol. 8, No. 2, pp. 95--128, 1994.

[101] Savoji, M.H. Endpointing of Speech Signals. Speech Communication, Vol. 8, No. 1, March 1989, pp.46-60

[102] Shen, X. N. Relative duration as a perceptual cue to stress in Mandarin, Language and Speech 36(4): 41-433, 1993.

[103] Shen, Xiao-Nan, Interplay of the four citation tones and intonation in Mandarin Chinese, in Journal of Chinese Linguistics, vol. 17, no. 1, pp. 61-74, 1989.

[104] Shen, Xiao-Nan. The Prosody of Mandarin Chinese. University of California Press, 1990.

[105] Shih, Chilin and Sproat, Richard. Issues in Text-to-Speech Conversion for Mandarin. Computational Linguistics and Chinese Language Processing, 1(1), 37-86, 1996.

[106] Shih, Chilin, and Kochanski, Greg. Chinese Tone Modeling with Stem-ML. In Proceedings of the International Conference on Spoken Language Processing, (Beijing, China), ICSLP, 2000.

[107] Shih, Chilin. The Prosodic Domain of Tone Sandhi in Mandarin Chinese. PhD Dissertation, UC San Diego. 1986.

[108] Shih, Chilin. Tone and Intonation in Mandarin. In N. Clements (ed). Working Papers of the Cornell Phonetics Laboratory, No. 3, 83-109. 1988.

[109] Silverman, K.; Beckman, M.; Pitrelli, J.; Ostendorf, M.; Wightman, C.; Price, P.; Pirerrehumbert, J., and Hirschberg. J. ToBI: A Standard for Labeling English Prosody. In Proceedings of ICSLP 92, Volume 2, pages 867-870,1992.

[110] Speechworks. Assessing text-to-speech system quality, http://www.tmaa.com/tts /Evaluating%20TTS%20Systems%20White%20Paper%2010-02.pdf, 2002.

[111] Speer, S. R., Shih, C.-L., & Slowiaczek, M. L. Prosodic structure in language comprehension: Evidence from tone sandhi in Mandarin. Language and Speech, 3 2,337-354. 1989.

[112] Sproat, R. and Olive, J. An Approach to Text-To-Speech Synthesis, in Speech Coding and Synthesis, pp. 611--633, Elsevier, 1995.

[113] Sproat, R., editor, Multilingual Text-to-Speech Synthesis: The Bell Labs Approach, Kluwer Academic Publishers, 1998.

[114] Sproat, R., Hirschberg J., and Yarowsky D. A Corpus-Based Synthesizer. International Conference on Spoken Language Systems, Banff, Canada, 1992, pp. 563-566.

[115] Sproat, Richard. Test Interpretation for TTS Synthesis. In Survey of the State of the Art in Human Language Technology, ed. Ron Cole, 1995.

[116] Sun, X. and Applebaum, T.H. Intonational Phrase Break Prediction Using Decision Tree and N-Gram Model, Proc. of 7th European Conference on Speech Communication and Technology (Eurospeech), Aalborg, Denmark, Vol 1, pp. 537-540, 2001.

[117] Taylor, P. A. 1995. The Rise/Fall/Connection Model of Intonation. Speech Communication, 15, 169--186.

[118] Taylor, P. and A.W. Blank. Assigning Phrase Breaks From Part-of-speech Sequences. Computer Speech and Language, 12:99-117,1998.

[119] Taylor, P. and Black, A. W. Synthesizing Conversational Intonation from a Linguistically Rich Input. In Proc. ESCA Workshop on Speech Synthesis, Mohowk, NY., 1994.

[120] Taylor, Paul A. "The Tilt Intonation Model", in ICSLP98, 1998.

[121] Taylor, Paul A. Synthesizing Intonation Using the Rise/Fall/Connection Model. In Proc. ESCA Workshop on Prosody, Lund, Sweden, 1993.

[122] Taylor, Paul A.; Black, Alan W. and Caley, Richard J. The Architecture of the Festival Speech Synthesis System. In Third International Workshop on Speech Synthesis, Sydney, Australia, November 1998.

[123] Traber, C. F0 generation with database of natural F0 patterns and with neural network. In G. Bailly, C. Benoit, and T. R. Sawallis, editors, Talking machines, Theories, models, designs. pp:287-304, Elseview Science, 1992.

[124] Veilleux, N., Ostendorf, M., Price, P., and Shattuck-Hufnagel, S. Markov Modeling Of Prosodic Phrase Structure. In Proc. Int. Conf. on Acoustics, Speech and Signal Processing, volume 2, pages 777--780, Albuquerque. 1990.

[125] Veilleux, N.M., M .Ostendorf, P. J. Price, and S. Shattuck Hufnagel. Markov modeling of prosodic phrase structure. In International Conference on Speech and Signal Processing. IEEE, 1990.

[126] Wang, Changfu, Fujisaki, H. Tomana, R. and Ohno, S. Analysis of fundamental frequency contours of standard Chinese in terms of the command-response model and its application to synthesis by rule of intonation. In proceedings of ICSLP, 2000.

[127] Wang, Changfu, Fujisaki, H., Ohno. S., Kodama, T. Analysis And Synthesis Of The Four Tones In Connected Speech Of Standard Chinese Based On A Command-Response Model, in Proc. ICSLP 2000, Beijing China, 2000.

[128] Wang, M.Q. and Hirschberg, J. Automatic Classification of Intonational Phrase Boundaries. Computer Speech and Language, 6: 175-196,1992.

[129] Wang, Ren Hua, Liu, Qing Feng, and Tang, Difei. A New Chinese Text-to-Speech System with High Naturalness. In Proceedings of the International Conference on Spoken Language Processing, (Philadelphia, USA), ICSLP, 1996.

[130] Wang, Ren Hua. Overview of Chinese Text-to-Speech Systems. Communications of Chinese and Oriental Languages Information Processing Society 8 (2), pp. 221-234, 1998.

[131] Wang, Ren-Hua, Ma, Zhongke. Li, Wei, and Zhu, Donglai, A Corpus-Based Chinese Speech Synthesis with Contextual-Dependent Unit Selection. In Proceedings of the International Conference on Spoken Language Processing, (Beijing, China), ICSLP, 2000.

[132] Wang, W.J., Campbell, W.N., Iwahashi, N., and Sagisaka, Y. Tree-Based Unit Selection for English Speech Synthesis, Proc. ICASSP'93, Minneapolis, Vol. 2, pp. 191-- 194. 1993.

[133] Wu, Chung-Hsien; Chen, Jau-Hung. Automatic generation of Synthesis Units and Prosodic Information for Chinese Concatenative Synthesis, Speech Communication, vol. 35, 219-237, 2001.

[134] Wu, J. R. Wang, Z. L. et al. Chinese-English dictionary. Commercial Printing House, Beijing, China, 1989.

[135] Xu, Yi. Contextual tonal variations in Mandarin, Journal of Phonetics, 25, 61-83. 1997.

[136] Xu, Yi. Effects of tone and focus on the formation and alignment of F0 contours, Journal of Phonetics, 27: 55-105, 1999.

[137] Yi, Jon. Natural sounding speech synthesis using variable-length units, Master's thesis. MIT, 1997.

[138] Yu, Shiwen, et al. The specification of Basic Processing of Contemporary Chinese Corpus. Journal of Chinese Information Processing, Issue 5 & 6. 2002.

Appendix

A. Part-of–speech Tag Set of Peking (Beijing) University

Tag	Chinese Name	Translation
Ag	形容词性语素	Adjective morpheme
a	形容词	Adjective
ad	副形词(直接作状语的形容词)	Adjective used as adverbial modifier
an	名形词(具有名词功能的形容词)	Active with noun function
b	区别词	Discriminate
c	连词	Conjunction
d	副词	Adverb
Dg	副语素	Adverb morpheme
e	叹词	Exclamation
f	方位词	Noun of locality
g	语素(大多能作为合成词的词根)	Morpheme
h	前接成分	Prefix
i	成语	Idiom
j	简称略语	Abbreviation
k	后接成分	Postfix
l	习用语	Idiom
m	数词	Numeric
Ng	名语素	Noun morpheme
n	名词	Noun
nr	人名	Personal name
ns	地名	Place name
nt	机构团体	Name of organ and party
nz	其他专名	Other proper noun
o	拟声词	Onomatopoeia
p	介词	Prepositional
q	量词	Quantity
r	代词	Pronoun
s	处所词	Space
Tg	时语素(时间词性语素)	Time morpheme
t	时间词	Noun of time
u	助词	Auxiliary
Vg	动语素(动词性语素)	Verb morpheme
vd	副动词(直接作状语的动词)	Adverb verb
vn	名动词(具有名词功能的动词)	Verb Noun
w	标点符号	Punctuation
x	非语素字(符号)	Symbol
y	语气词	Modal
z	状态词	Adjective of state

B. Features for Unit in Speech Inventory

Feature	Description	Type	Range	Remarks
CurrInit	Initial of the syllable	Category	1-22	
CurrFinal	Final of the syllable	Category	1-38	
CurrTone	Tone of the syllable	Category	1-5	
BreakLeft	Break type before the syllable	Category	0-4	
BreakRight	Break type after the syllable	Category	0-4	
PrevInit	Initial of the previous syllable	Category	0-22	0 for no previous syllable
PrevFinal	Final of the previous syllable	Category	0-38	0 for no previous syllable
PrevTone	Tone of the previous syllable	Category	0-5	0 for no previous syllable
NextInit	Initial of the next syllable	Category	0-22	0 for no next syllable
NextFinal	Final of the next syllable	Category	0-38	0 for no next syllable
NextTone	Tone of the next syllable	Category	0-5	0 for no next syllable
Duration	Duration of the syllable	float	float	
EnergyRMS	Energy of the syllable	float	float	
PitchMean	Pitch mean of the syllable	float	float	
PitchStart	Pitch value of the start point of the voiced part	float	float	
PitchMiddle	Pitch value of the middle point of the voiced part	float	float	
PitchEnd	Pitch value of the end point of the voiced part	float	float	
PitchRange	Pitch range of the syllable.	float	float	
EnergyHalfPoint	Percentage position of ½ energy dividing.	float	[0,1]	
EnergyStart	RMS Energy of start point of syllable.	float	float	
EnergyEnd	RMS Energy of end point of syllable.	float	float	

C. Sentences for Listening Testing

1. 超负荷的工作累倒了王柏林
2. 承包或租赁转让金收不回来
3. 反映了周恩来作为开国总理
4. 每年节约经费二百余万元
5. 那么妇女状况也难以改善
6. 敲击电脑键盘声不绝于耳
7. 陆军参谋长和外长进行磋商
8. 在天安门城楼的灯笼里
9. 中国和美国由于文化原因
10. 此案案发五年多的时间
11. 凡单位一次购车五辆以上的
12. 就要写到东北解放战争
13. 一个人独立完成证券的交易
14. 民族医药业应采取积极对策
15. 一位日本人突然找到我家
16. 这意味着用于满足人们学习
17. 坐落在南京路西藏路口
18. 变要我服务为我要服务
19. 并为其注入实质内容
20. 还为人们提供了高倍望远镜
21. 荒漠丛林中奋勇跋涉的脚步
22. 加快内引外联的步伐
23. 教育科学文化卫生委员会
24. 平均每月为七百三十六元
25. 葡萄牙经过数年的艰苦努力
26. 三年两载可能还成不了形
27. 她拉着我大步进了楼又说道
28. 放映室的灯光亮了
29. 王秀英摄于坦桑尼亚
30. 才能凝成这泥土的精华
31. 单等对方安排职工来听课
32. 冷冻货源源送往港澳市场
33. 熊熊烈焰映红了大半个天空
34. 音乐剧要求演员歌舞戏全能
35. 澳门增加委员名额问题
36. 关于堡贸易政策问题
37. 收费标准低于航空包裹资费
38. 营造有利于开展革命传统
39. 改革前后的场景接续起来
40. 工商部门优先办理营业执照
41. 实践和胜利的二十年
42. 是心胸博大有力量的国家
43. 收拾完卷宗刚要回家
44. 维护文明环境需要众人齐努力
45. 伟大的朋友影片摄成
46. 专门用于奖励热爱新闻事业
47. 北京西藏大厦一片欢歌笑语
48. 可溶性纤维就像小海绵一样

49. 马路两边顾客摩肩接踵
50. 门诊病人两天不能看病用药
51. 四川射洪县农村卫生见闻
52. 她因腿伤挥泪告别舞台后
53. 一些问题也随之暴露出来
54. 增设了灯光音乐喷泉
55. 创下我国农业最高劳动生产率
56. 给了我生命的欢悦与责任
57. 精神损失费若干了事
58. 乌克兰前外交部长乌多文科
59. 与国家骨干信息网络联通
60. 原子能部长米哈伊洛夫
61. 在人民日报实现了激光照排
62. 赞扬此次外交努力的成功
63. 这次轮训邀请了国防大学
64. 二月一日那天恰是正月初五
65. 当热气腾腾的饺子端上桌时
66. 老人还特意拿出节目单
67. 目前他已八十九岁高龄
68. 娘的一抹微笑一句夸奖
69. 使文明特色家庭成批涌现
70. 望着那依山傍水一望无边
71. 为打击仿冒美元纸币
72. 也不超越于客观实际
73. 帮助农民建设文化园地
74. 朝阳区团委为下岗职工献爱心
75. 大概还影响了若干文艺作品
76. 但罗马尼亚人似乎更老到
77. 但没有发生人员伤亡
78. 九十年代小说的现实主义精神
79. 她任中共湖南省工委秘书长
80. 因为有个主语更加明确一些
81. 又兼顾了与现行利率政策
82. 增强纳税人自觉纳税意识
83. 不要忘了给某号猪减料
84. 长野冬奥会闭幕之日
85. 共引种堡植物五百多种
86. 克林顿向美华人华侨贺春节
87. 李仍光舍身救人获金英勇勋章
88. 那旅游业还能蓬勃发展吗
89. 屈原闯荡天下尔后来归
90. 它们呆的温泉冒着热气
91. 埃斯特拉达已稳操胜券
92. 把自己的命运融入国家改革
93. 部队官兵每扫清一块雷区
94. 牡丹江市百万亩荒山披绿装
95. 任务指标虽然年年完成
96. 这样做符合美中两国利益
97. 作出了大力振兴电子工业
98. 六月六日是国际爱眼日
99. 可线条却像花岗石划过的
100. 来纪念馆参观有两个原因

D. Text Example for Intelligibility Testing

厚旺船皑额	龋径嫁南林	白日女最过	彝镰劣个职
隙用法裹好	腕样本摄你	狠隋威常囚	采前倪肩幅
阿总挝均韦	映达费雅蚀	落哀优年外	珊同诗条楔
肚解缘饮游	姿秆哗蜘谣	让国点且印	户崖腰纷初
茵粮场德电	矫版鲜辫大	颂爵古收更	快政并划而
妆神冬疆固	谦诬安笑火	管吟借爷睦	下坛打梅和
蹈头舅祁今	丫约海曾瑟	九悼两拄岂	爱小特儒柯
剪淹世尚亮	编兜玩捌口	使见颜补真	哈怎幸吃手
类趾要这墙	淆瘦莹恩人	舀戚致高烽	躯傲扮队然
商表北次演	号拖工禾速	诸学村急充	弯饿幢婉死
磁拉带况中	说但浇私五	博者富螟多	樱雀无影蚊
化除积软别	伟告李瘟匀	祸此辉萍全	等喂越励俞
没耳题乡回	务悸哪经柞	民每三体水	篱佑规钠卒
辱咳站航骗	有老少阳一	盏褥靶漳枉	拿烬距通景
索藕方触扳	走破鄙面才	饭塌二包泳	迄内蔬狰能
暗染刃亡发	仅添孩星当	宏羡渴妨地	从殃夜币桥
匙四习雍克	甥许门玛革	析六锣韵构	享肿哥泰盛
揩稳晴自狱	艳保抬牌蘑	车趣问竖桔	崇秽佳销仰
碌予洼策订	种忽象我备	琉滓需娃绅	看来非择舟
浓阻蚁阶官	新抱灶远在	青叫铲义亚	律摆胡熬干

E. List of Published Papers

1. **An Example-Based Approach For Prosody Generation In Chinese Speech Synthesis**, Dong Minghui, Lua Kim-Teng, International Symposium on Chinese Spoken Language Processing (ISCSLP 2000), Beijing, China, 2000.

2. **Using Prosody Database in Chinese Speech Synthesis**, Dong Minghui, Lua Kim-Teng, International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China, 2000.

3. **Prosodic Phrase Detection For Chinese TTS Using CART And Statistical Model**, Dong Minghui, Lua Kim-Teng, International Symposium on Chinese Spoken Language Processing (ISCSLP 2002), Taipei, 2002.

4. **Automatic Prosodic Break Labeling For Mandarin Chinese Speech Data**, Dong Minghui, Lua Kim-Teng, International Conference on Spoken Language Processing (ICSLP 2002), Denver, USA, 2002.

5. **Pitch Contour Model for Chinese Text-To-Speech Using CART and Statistical Model**, Dong Minghui, Lua Kim-Teng, International Conference on Spoken Language Processing (ICSLP 2002), Denver, USA, 2002.