Learning Prosodic Patterns for Mandarin Speech Synthesis

Yiqiang Chen Wen Gao
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, China 100080
Email: yqchen@ict.ac.cn

Tingshao Zhu
Dept. of Computing Science
University of Alberta
Edmonton, Canada T6G 2H1
Email: tszhu@cs.ualberta.ca

ABSTRACT

Higher quality synthesized speech is required for widespread use of text-to-speech (TTS) technology, and prosodic pattern is the key feature that makes synthetic speech sound unnatural and monotonous, which mainly describes the variation of pitch. The rules that are now being used in most Chinese TTS systems are constructed by experts, qualitatively and with low precision. In this paper, we propose a combination of clustering and machine learning techniques to extract prosodic patterns from actual large mandarin speech database to improve the naturalness and intelligibility of synthesized speech. Typical prosody models are found by clustering analysis, some machine learning techniques including Rough Set, ANN and Decision tree are trained respectively for fundamental frequency and energy contours, which can be directly used in a pitch-synchronous-overlap-add-based (PSOLA-based) TTS system. The experimental results showed that synthesized prosodic features quite resembled their original counterparts for most syllables.

Keywords:

TTS, Pitch, Mandarin Speech Synthesis, Data Mining

1.INTRODUCTION

Text-To-Speech (TTS) technology is currently useful only in a limited number of applications because the quality of synthetic speech is not as good as people expected. Prosody, which includes the phrase and accent structure of speech, is one of important component for TTS system. In the field of speech signal process, pitch (fundamental frequency and F0) is the most mysteriously expressive of the prosodic phenomena, and the variation of pitch in speech can be used to express the speaker's intention, especially in Mandarin.

Although many researchers have proposed some prosodic variation patterns, the patterns are described qualitatively, or with great limitation. Wu [1][2] found that in Mandarin when syllables are combined, their tones changed to be continuous, and he gave some qualitative rules. Chu [3] uses some pitch patterns in Chinese speech synthesis system, including 14 kinds of pitch shapes of isolate syllables and 22 kinds of shapes of two-word phrases, but obviously only these shapes can't describe the variations of Mandarin to a large extent.

In recent years, some researchers intend to learn the variation patterns base on large speech database. Lee S. and Oh Y-H [4] describes the tree-based modeling of prosodic phrasing, pause duration for Korean TTS system. Ostendorf [5] describes a dynamical system model for generating fundamental frequency, which allows automatic estimation of parameter from labeled large speech database. Hu [6] proposed a template-driven generation of prosodic information for Chinese text-to-speech conversion. Ross KN. Chen [7] proposed a new RNN-based prosodic information synthesizer for Mandarin Chinese text-to-speech. Cai [8] establish a Chinese text to speech system and a prosody learning system based on NN.

Although these methods have made advances, they are still far away from reaching the goal of generating proper prosodic information for synthesizing speech with high naturalness. The drawback lies in their inability to elegantly invoke higher-level linguistic features in exploring the prosodic phrase structure of Mandarin speech. This motivates us to use a combination of clustering and ML techniques to learn prosodic variation patterns to improve the naturalness and intelligibility.

This paper is organized as follows. Sections 2 introduce TTS, Sections 3 discuss the clustering of prosodic pattern. The Training process and the prediction are described in Section 4, and some conclusions of our on-going research will be given in Section 5.

2. TEXT TO SPEECH

The three main steps in the TTS process are illustrated in Figure.1.

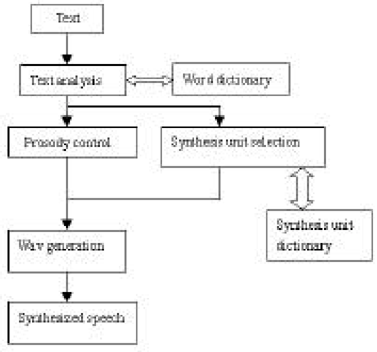

Figure 1 TTS system

Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000
(O.R. Zaïane, S.J. Simoff, eds.)

Speech synthesized from text suffers from a major shortcoming - it sounds synthetic. It is found that synthetic speech significantly more difficult to comprehend than natural speech. There is broad consensus that poor prosody is the key feature that makes synthetic speech sound unnatural and monotonous. Prosody conveys the relative importance of words by making these words stand out acoustically (prominence), and helps listeners "chunk" the message by inserting acoustic punctuation marks (phrasing) as pauses, decreases in rate, and characteristic pitch movements. Speech with inappropriate prosody prevents listeners from understanding the spoken message. It will also undermine the credibility and effectiveness of the animated character that produces it.

Prosody in TTS systems involves three levels. First, text analysis components compute phrase boundary locations and prominence ("prosodic structure"). Second, acoustic prosodic components compute phoneme duration, fundamental frequency (or pitch) contours, and (optionally) contours for additional acoustic parameters such as amplitude or spectral tilt. Finally, signal processing components compute a digital speech wave that expresses the phoneme sequence having the desired timing and pitch contour.

The main goal of our work is to improve the prosodic structure of speech generated from text. The TTS system must analyze text and use this information to generate both a symbolic structure (e.g., locations and type of phrase boundaries and prominence of important words) and an acoustic waveform that expresses the desired meaning of each utterance.

This general problem of lack of an appropriate prosodic model was encountered in Mandarin TTS prosodic information synthesis. Mandarin Chinese is a tonal language. Each character is pronounced as a syllable. Only about 1300 phonetically distinguishable syllables comprise the set of all legal combination of 411 base-syllables and five tones. Each base-syllable is composed of an optional consonant initial and a vowel final. The word, which is the smallest syntactically meaningful unit, consists of one to several syllables. Because syllables are the basic pronunciation units in Mandarin speech, they are commonly chosen as the basic synthesis units in Mandarin TTS systems. Accordingly, the prosodic information that must be synthesized includes syllable pitch (F0) contour, syllable energy contour, syllable initial and final duration, as well as intersyllable pause duration. Among them, syllable pitch contour has the most important effect on naturalness of synthetic speech. So pitch contour synthesis is of primary concern in Mandarin TTS. The tone of syllable is mainly determined by its F0 contour. However, the pronunciation is usually highly context-dependent, It would seem that syllable F0 contour are various modification in continuous speech. Therefore, the F0 generation is not a trivial task. There are many methods have been proposed in the past, including rule-based methods [9][10], statistical model-based methods [4][6], and MLP-based methods [7]. In contrast to other researchers who employ a single technique, we present a combination of clustering and ML techniques, applied for the identification of relationship(s) between sound characteristics of speech pattern and linguistic features of the corresponding text.

We assume that the F0 contour in the continuos speech data are not variety randomly but can be obtained through modifying some classic F0 model with duration and mean. These classic F0 models can be obtained from the preprocessed actual F0 contours.

3.DATA PROCESSING

3.1 Speech Database

The speech corpora and the labeled speech corpora will be used for our acoustic prosody and signal processing efforts. The Speech Database that we are using is a Chinese speech synthesis database called CoSS-1. CoSS-1 includes the pronunciation of all isolate syllables, the 2-4 word phrases and some sentences. The number of isolate syllables with tone is 1268, and that of word phrase is 1640 and sentence is 210.

CoSS-1 records the speech wave and laryngograph synchronously. The sampling rate is 16000/s, and each sample is stored in two bytes. The sentence in the database covers almost the whole tone collocations in Chinese pronunciation.

The preprocessing mainly deals with the data from speech database directly, which extracts pitch, wraps the duration and normalizes and smooth and zero mean the pitch values to meet the requirement of cluster algorithm.

3.2 Pitch Extraction

To learn the patterns, the pitch should be calculated at first. There are many methods to extract pitch from speech wave, but the precision is very low [11]. Since we want to learn the patterns and use them to generate pitch after training, the accuracy is very important.

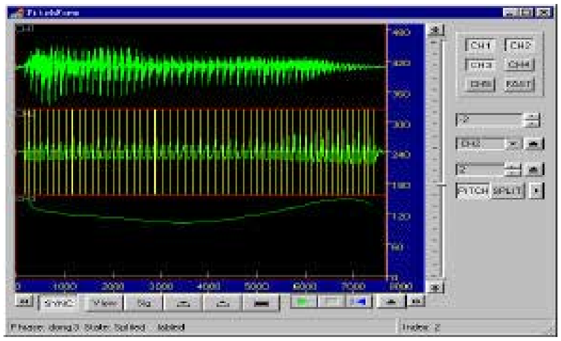

Figure. 2. A snapshot of Pitcher.

A tool called Pitcher is implemented to extract pitch from laryngograph. It works by annotating each cycle's beginning and ending point, then calculating the pitch. Let X_i be the beginning point of one cycle and X_j be the ending point, then the pitch of this cycle should be $16000/(X_j - X_i)$. Pitcher can also be used to split phrases and play the speech data. Figure 2 gives a snapshot of Pitcher.

Pitcher can be used to split a phrase and annotate each syllable.

The results of splitting and annotating are stored in database, then the algorithm can retrieve them for training and testing. To annotate pitches, you should firstly sign the reference cycle, then the beginning and ending point of the period. Pitcher deals with cycles one by one within the period to calculate pitches.

3.3 Time Wrapping and Normalization

The length of pitches that should acts as the training examples differs from each other significantly. A new algorithm is designed to wrap the pitches, which differs from the traditional time wrapping method DTW [11](Dynamic Time Wrapping) which is widely used in speech signal process. Figure 3 gives the new algorithm.

For the speech data we used, the pitches' value domain is between $50-260$. In this paper, the following equation is used to normalize the pitch value.

$$Normalized = (Pitch - min) / (max - min) \qquad (1)$$

Where *max* is the maximum of all pitches' value and *min* is the minimum. *Pitch* stores the pitch to be calculated and *Normalized* is the normalized value.

3.4 Smoothing and filter: There are many filter algorithms in signal processing [11]. In this paper, the window design filter is presented to eliminating the large fluctuant data. Assume a sequence of observations $X=[x_0,x_1,\dots,x_n]$, the windows width is m. we can gain another sequence states $Y=[y_0,y_1,\dots,y_{n-m+1}]$ with the following formulation (2):

$$y_i = \sum_{j=i}^{i+m} x_j \Big/ m \qquad (2)$$

3.5 Zero-mean: To avoid the effect of F0 energy, zero-mean method is proposed for each pitch fundamental frequency. The zero-mean method represents a sequence of observation $X=[x_0,x_1,\dots,x_n]$ in terms of a sequence of states $Y=[y_0,y_1,\dots,y_n]$ with the following equation (3):

$$y_i = x_i - \sum_{j=1}^{n} x_j \Big/ n \qquad (3)$$

3.6Clustering method: The quantification of the similarity notion is important for clustering, and in our clustering, we use the following one:

$$Dist(X,Y) = \sqrt{\sum_{i=1}^{n}((x_i-\bar{x})-(y_i-\bar{y}))^2} + \left|\bar{x}-\bar{y}\right| \qquad (4)$$

Where $X=(x_1,x_2,\dots,x_n)$, $Y=(y_1,y_2,\dots,y_n)$, $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ and $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$. This method can calculate the similar more precise than the Euclidean distance method.

The ISODATA [12](Iterative Self-organizing Data) algorithm is chosen for our clustering, the main procedures are the following:

3.6.1Present the clustering parameters

C: number of expected classes; *MaxIterate*: the Max times for adjusting; ***MinSamples:*** the Min number of objects in one class; ***I****:* combination parameter; ***J:*** partition parameter.

3.6.2Choose initial cluster centers

Calculate the mean $\overline{x_i}$ and variance $s_i(i=1,2,3,\dots,n)$

Arbitrarily choose 2n+1 objects as the initial clustering centers ： $\overline{X}=\left(\overline{x_1},\overline{x_2},\overline{x_3},\dots,\overline{x_n}\right)$ and $\left(\overline{x_1},\overline{x_2},\dots,\overline{x_i}\pm s_i,\dots,\overline{x_n}\right), i=1,2,\dots,n,$

3.6.3Classify and adjust the objects based on K-means algorithm

If there is no re-distribution of the objects in any cluster happens or the max times ***MaxIterate*** for adjusting is achieved, the process terminates, otherwise, the adjusting will be repeated as following:

Deleting*：** if the number of objects in some class is less than ***MinSample, then the class should be deleted, at the same time, the objects in that class will not be reused.

***Partition*：** assume that m classes are generated after several times overlapping, and there must be one character in n of each class holding the Max variance. Let

$$S_{threshold} = \overline{s_{\max}} \bullet \frac{J}{1+e^{-(m-C)}}$$

Where $\overline{s_{\max}}$ reprsent the mean of max variance of all the classes.

To each class, the max variance S_I of every character can be calculated，if S_I>$S_{threshold}$，then this class should be partitioned as following: $\left(\overline{x_1},\overline{x_2},\dots,\overline{x_i}\pm s_i,\dots,\overline{x_n}\right), i=1,2,\dots,n$

Combination: assume that m classes are generated after several times overlapping, and the min distance value between every two centers can be obtained. Let

$$D_{threshold} = \overline{D_{\min}} \bullet \frac{I}{1+e^{-(m-C)}}$$

Where $\overline{D_{\min}}$ reprsent the mean of min distance of all the classes.

To every two classes, if the distance between their centers is less than $D_{threshold}$, then they are combined, and the center of new class should be recalculated. After clustering, there are 18 F0 pattern are classified, Figure 3 shows them:

After clustering, there are 18 F0 pattern are classified, Figure 3 shows them:

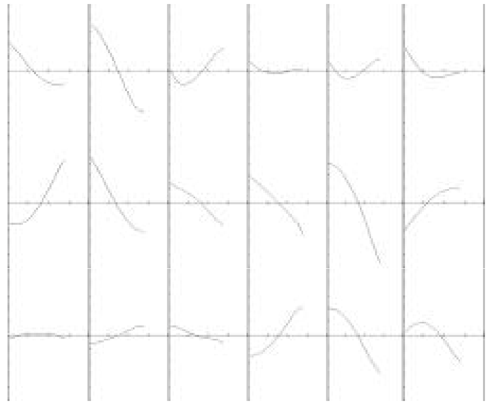

Figure 3: F0 patterns after clustering analysis.

The original pitch from sentences is discreted with extracted classic F0 models, and at the same time the original length and mean should be kept for future learning. The original prosody pattern is preprocessed into three parts: zero-mean F0 pattern, duration, and mean.

4. PROSODIC LEARNING

4.1 Linguistic Features

Our aim is to explore the relationship between the prosodic pattern of Mandarin speech and the linguistic features of the input text to simulate human's prosody pronunciation mechanism.

The Chinese Dictionary that we are using includes the spell, vowel, constant, tone, part of speech (POS), and some word syntax and semantic. From this dictionary and the existence of a text processing model, the lexical information (phonemic representations and lexical stress) and symbolic prosodic markers can be obtained. In this paper, after parsing, the linguistic features including the following:

The number of pitch in word
The sequence number serial number of pitch in word
Word class and POS
Is substantive or function word?
Is prediction or noun word?
The vowel, constant and tone of current pitch
The vowel, constant and tone of prior and post pitch

4.2 Feature Selection

Before training, the Rough set [13][14] is proposed to find the minimum attribute set. The rough set theory is based on indiscernibility relation. Suppose four finite, non empty sets R, A, V and f, where R is the universe, and A is a set of attributes, V is the value set of each attribute and f is a function map $f(U, A) \to V$. The indiscernible relation I is associated with every subset of attributes $P \in A$ and defines as: $I(P) = \{(r_i, r_j) \in U \times U : f(r_i, attr) = f(r_j, attr), \forall attr \in P\}$

Where $f\ (r_i,\ attr)$ is the value of attribute *attr* in object r_i. If $(r_i, r_j) \in I(P)$, then r_i and r_j are P-indiscernible.

Rough set can remove unnecessary attributes from the set A by considering redundancies and dependencies between attributes. Let P be a subset of A, and the initial P is the set A. If $I(P) \neq I(P - \{attr\})$, then we say that the *attr* can be moved from the set A. Thus the main features are selected by Rough set. The main features are used as input of ANN and the condition attributes of decision tree. We construct three ANN or decision trees respectively, they can predict the F0 model, the F0 mean and the F0 duration.

4.3 Training and Prediction

There are many kinds of neural networks, which can be used for learning. We intend to learn the mapping between the linguistic features and the F0 mean value. Since backpropagation network has implicit input layer and output layer [15], and it can also give very good result, thus it is chosen to be trained in our system.

In order to generate training and testing data, all the sentences are split firstly, calculating the pitches, wrapping the pitches to the same length, normalizing pitches' value and discrete the pitch. Then the pitch class, the linguistic parameters obtained by text parsing are labeled for neural net training and testing. For the network learning the F0 model, its input layer consists of 28 units, and the hidden layer consists of 34 units. There is only one unit in output layer. The input layer's units are described as Table 1.

Table 1: the definition of input layer

Number of units	Description
4	Length of word(1-4)
4	Pitch's location in word
5	Part of speech
6	Vowel/consonant
3	Tone of pitch
3	Tone of previous pitch
3	Tone of next pitch

The training of the F0 length and the F0 mean are as same as the training of F0 model. Then Three different neural networks are constructed to predict the F0 model, the F0 mean and the F0 length respectively.

Using the same linguistic parameters as condition attributes and the F0 model, the F0 mean and the F0 lengths as decision attribute, three different decision trees are constructed respectively. The C4.5 system [16], which has many advantages in building decision tree, is selected for our construction.

After training, the NN and the decision tree can be used to predict and generate the fundamental frequency. We compared two ways and results shows in Table 2.

OL means original Length of pitch
LPDT means Length predicted by decision tree
LPNN means Length predicted by ANN
OM means original mean of pitch
MPDT means mean predicted by decision tree
MPNN means mean predicted by ANN

Table 2: some predict result of NN and Decision tree

OL	LPDT	LPNN	OM	MPDT	MPNN
(zhi2) 24	17.5	27.8	261.4	185	246.4
(pai2) 45	37.5	29.1	234.4	175	239.2
(shi4) 32	27.5	36.2	251	265	235
(ran2)38	7.5	31.5	197	165	167.5
(qi4) 26	27.5	24.2	275.5	285	248.5
(re4) 35	32.5	24.1	277.3	255	255.3
(shui3)5	17.5	8.9	173	245	163.8
(qi4)26	27.5	24.4	211.6	185	206.7
(yan2)31	22.5	21.8	196	215	198
(jin4) 16	27.5	22.2	252.7	235	249.8
(an1) 24	17.5	32.2	249.5	315	239.3
(zhuang4)36	17.5	41.4	275.9	275	262.4
(zai4)28	42.5	30.1	233.3	305	277.3
(yu4)43	27.5	34	266	285	257.2
(shi4)6	37.5	25.5	173.7	185	189
(nei4)30	22.5	23.2	238.9	305	238.6
(shi3)12	27.5	14.7	140.9	185	144.2
(yong4)10	37.5	20.3	168.6	245	195.1

The experiment was taken on our labeled large speech database, the variation of the data between the original data and the predicted one was calculated as follow:

$$mean = \frac{\sum_{i=1}^{n} (prediction - original)}{n}$$

$$variation1 = \frac{\sum_{i=1}^{n} (prediction - original)^2}{n}$$

$$variation2 = \frac{\sum_{i=1}^{n} (prediction - original - mean)^2}{n}$$

$$unsimilarity = \frac{\sqrt{variation1} + \sqrt{variation2} + mean}{3}$$

Table3 shows the variations of predicted result:

Table 3: The variations of predicted result

variation	F0 model	F0 duration	F0 mean
Decision tree	2.6	5.9	33.8
ANN	1.8	3.3	11.7

The results show that the decision tree is not good at predicting the continuous attribute while ANN can do it well. Thus the decision tree is only be used to predict the F0 model while the BP was used to predict the F0 mean and F0 length. The linguistic features are reselected focus on each prediction respectively, and Table 3,4 shows the summary of our features selection:

Table4: Attributes of Decision tree for F0 model

Condition Attributes	Number of pitches in word (len)
	Series number of pitches(wordno)
	Part of Speech (type)
	Substantive or function word (xs)
	prediction or noun word(tw)
	Current tone, pretone, posttone
Decision	F0 model

Some rules for F0 model prediction：
type = 1 and tone = 2 and pretone = 3 and posttone = 2-> class 4
len = 3 and type = 1 and tone = 2 and posttone = 5 -> class 4
type = 4 and tone = 2 and pretone = 5 and posttone = 4 -> class 13
type = 6 and tone = 2 and pretone = 4 -> class 14

Table 5: the definition of input/output layer

Input layer Definition	Number of pitches in word (len)
	Series number of pitches(wordno)
	Part of Speech (type)
	Substantive or function word (xs)
	Prediction or noun word(tw)
	Consonant and tone (pycon, tone)
	pretone, posttone
Output of NN	Length of F0(discreted)

4.4Experiment

The F0 models are predicted by decision tree while the F0 duration and mean are predicted by ANN. The F0 model can be modified in accordance with F0 duration and F0 mean. The modification is based on a simple interpolation method. The experiment was taken on the CoSS-1 speech database, some experiment results are showed in figure 4.

zhi pai shi re shui qi yan jin an zhuang zai yu shi nei shi yong

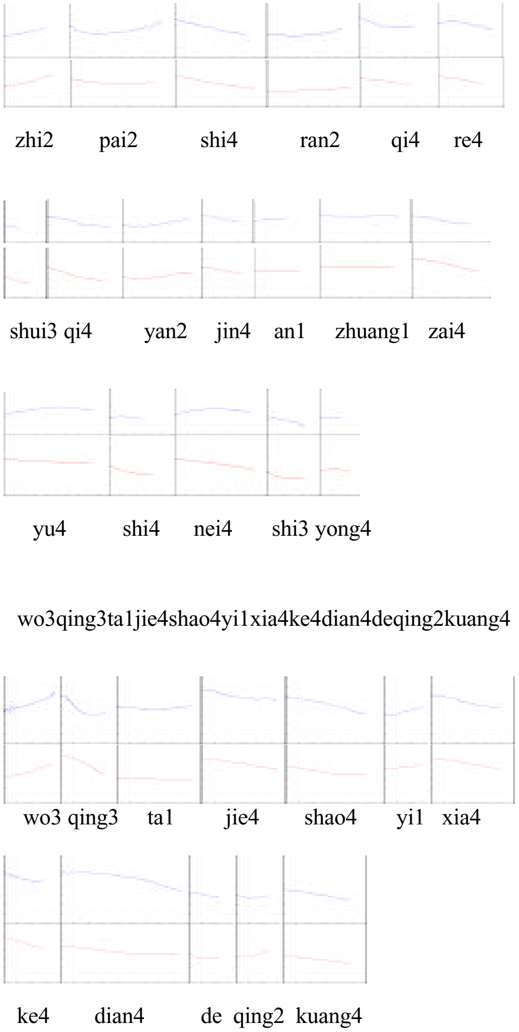

Figure 4 above is original pitch, lower is synthesis one

5.CONCLUSION

In this paper, we propose a combination of clustering and machine learning techniques to extract prosodic patterns from actual large mandarin speech database to improve the naturalness and intelligibility of synthesized speech. Typical prosody models are found by clustering analysis, some ML techniques including Rough Set, ANN and Decision tree are trained respectively for fundamental frequency and energy contours, which can be directly used in a pitch-synchronous-overlap-add-based (PSOLA-based) TTS system. The prediction result of ANN and Decision Tree can be combined to generate the fundamental frequency and energy contours. So, the effects of high-level linguistic features on prosodic information generation are well handled. The experimental results showed that synthesized prosodic features quite resembled their original counterparts for most syllables.

6.REFERENCES

[1] Zongji Wu, "The tone variation in mandarin", *Chinese grammar*. No. 6, pp.439-449, 1982.

[2] Zongji Wu, "The design of prosodic rule for improving the naturalness of the Marian TTS", *The research on Chinese language and words*, *Tsinghua University press,* pp.355-365, 1996.

[3] Min Chu, "Research on Chinese TTS system with high intelligibility and naturalness", *Ph.D thesis*, Institute of Acoustics, Academia Sinica, 1995.

[4] Lee S, Oh Y-H, "Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS system", *Speech Communication*, Vol.28, No.4, pp.283-300, 1999.

[5] Ross KN, Ostendorf M, "A dynamical system model for generating fundamental frequency for speech synthesis", IEEE *Transaction on speech and audio processing*, Vol. 7, No. 3, pp.295-309, 1999.

[6] Chung-Hsien Hu, Jan-Hung Chen, "Template-driven generation of prosodic information for Chinese concatenate synthesis", IEEE *International Conference on Acoustics, Speech, and Signal Processing*, Vol.1, pp.65-68, 1999.

[7] Sin-Horng Chen, Shaw-Hwa Huang, Yih-Ru Wang, "An RNN-Based Prosodic Information Synthesizer for Mandarin Text-to-Speech", IEEE *Transaction on speech and audio processing*, Vol. 6, No. 3, pp.226-239, 1998.

[8] Cai Lianhong, Zhang Wei, Hu Qiwei, "Prosody learning and simulation for Chinese text to speech system", *Qinghua Daxue Xuebao/Journal of Tsinghua University*, Vol.38, No.S1, pp.92-95, 1998.

[9] L.S.Lee, C.Y. Tseng, and M. Ouh-Young, "The synthesis rules in a chinese text-to-speech system", *IEEE trans. Acoust., speech, signal Processing,* Vol. 37, pp. 1309-1320, 1989.

[10] C.H. Wu, C. H. Chen, and S. C. Juang, "An CELP-based prosodic information modification and generation of Mandarin text-to-speech", *in proc. ROCLING VIII*, pp. 233-251, 1995.

[11] L.Rabiner and B.Juang. "Fundamentals of Speech Recognition." *TsingHua University Publishing Company*. 1999.

[12] Bian Zhaoqi and Zhang Xuegong, "Pattern recognition", *TsingHua University Publishing Company*. 1999.

[13] Pawlak Z, "Rough classification", *International Journal of Human-Computer studies*, Vol.51, No.2, pp.369-383, 1999.

[14] Walczak B, Massart DL, "Rough sets theory", *Chemometrics &Intelligent Laboratory Systems*, Vol.47, No.1, pp.1-16, 1999.

[15] Wang Wei. Principle of Artificial Neural Network ---- rudiment and implement. *Beijing University of Aeronautics and Astronautics Press*, 1995

[16] J.Ross Quinlan, "C4.5: Programs for Machine Learning", *Morgan Kaufmann Publishers press*, 1993.