ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Multi-Strategy Data Mining on Mandarin Prosodic Patterns

Yiqiang Chen[1], Wen Gao[1], Tingshao Zhu[2], Jiyong Ma[1]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China 100080
Email: {yqchen, wgao, jyma}@ict.ac.cn
[2]Dept. of Computing Science, University of Alberta, Edmonton, Canada T6G 2H1
Email: tszhu@cs.ualberta.ca

ABSTRACT

Mandarin prosodic models are very important in speech research and synthesis, which mainly describes the variation of pitch. The models that are now being used in most Chinese Text-To-Speech systems are constructed by expert, qualitatively and with low precision. In this paper, we propose a Multi-strategy Data Mining framework to extract prosodic patterns from actual large Mandarin speech database to improve the naturalness and intelligibility of synthesized speech. In data preprocessing, typical prosody models are found by clustering analysis, and Rough Set is employed for feature selection. ANN and Decision tree are trained respectively. The prediction result of ANN and Decision Tree are integrated to generate fundamental frequency and energy contours. The experimental results showed that synthesized prosodic features quite resembled their original counterparts for most syllables.

1. INTRODUCTION

Text-To-Speech (TTS) technology is not widespread used because of the low quality. Prosody, which includes the phrase and accent structure of speech, is one of important component for TTS system. Although many researchers have proposed some prosodic variation patterns, the patterns are described qualitatively [1][2], and thus cannot be used in speech synthesis directly.

In recent years, some researchers intend to learn the variation patterns base on large speech database. Following this way, it is possible to extract patterns in quantity, and they can be used in speech synthesis directly to improve the quality of synthesized speech. Lee S. And Oh Y-H [3] describes the tree-based modeling of prosodic phrasing, pause duration for Korean TTS system. Ross KN, Ostendorf M [4] describes a dynamical system model for generating fundamental frequency, which allows automatic estimation of parameter from labeled large speech database. Chun-Hsien Hu[5]proposed a template-driven generation of prosodic information for Chinese text-to-speech conversion. Sin-Horng Chen [6] proposed a new RNN-based prosodic information synthesizer for Mandarin Chinese text-to-speech. Cai Lianhong [7] 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 natural-sounding speech.

Knowledge Discovery in Database (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understanding patterns in data [8]. It employs statistical and computational technology to extract useful patterns from large database [9]. KDD can also be called data mining. In this paper, multi-strategy data mining framework was proposed for generating prosodic information more precisely that can be incorporated into existing TTS synthesis system to improve the naturalness and intelligibility.

In data preprocessing, some typical prosody models are found by clustering analysis. Then these clusters and linguistic feature including tone combination, word length, part-of-speech (POS) of the word and word position in phrase obtained by text parsing are used as training data. The Rough set method is employed for feature selection. ANN (Artificial Neural Network) and Decision Tree are trained respectively using these features from a large labeled speech database. The prediction result of ANN and Decision Tree can be combined to generate the fundamental frequency and energy contours.

This paper is organized as follows. Sections 2 introduce our multi- strategy data-mining framework for prosodic information learning. Sections 3 discuss the classification of prosodic pattern. The Training process and the prediction are described in Section 4, and some conclusion of our on-going research will be given in Section 5.

2. MINING FRAMEWORK

The learning process is showed in Figure 1.

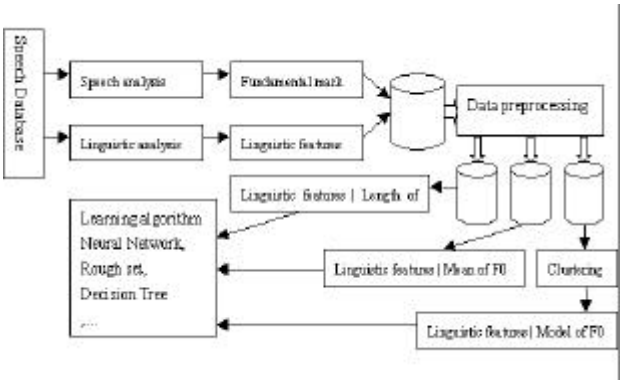

Figure 1:Multi-Strategy Mining framework

The linguistic analysis process is first presented, the existence of a text processing module can provide lexical information (phonemic representation and lexical stress) and symbolic prosodic markers. The F0 sequence of each pitch in a sentence can be obtained by fundamental marking. For each sequence, length normalizing, data smoothing and data filing are used for

preprocessing. The zero-mean sequence of each pitch can be gained after calculating the mean value, and each sequence is classified as one of these classical pitch models. Thus, the original sequence has been transformed to three parts: the F0 model, the length of original F0, and the mean of the normalized fundamental sequence.

The framework integrates some learning algorithm including Decision Tree, ANN, Rough set to learn the variance patterns of the three parts respectively. F0 contour is generated from the prediction of Decision tree and ANN. In our framework, we assume that the F0 contour can be generated from some typical F0 models by modifying duration and mean. The advantage of this model is that it can not only be automatic trained as templates model, but also facilitate using linguistic knowledge.

3. THE CLASSIC PROSODIC MODEL

We assume that the F0 contour in the continuous 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.

The classification 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.

The ISODATA [10](Iterative Self-organizing Data) algorithm is chosen for our clustering, the main procedures are the following:

1. Present 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.

2. Choose initial cluster centers
Calculate the mean $\overline{x_i}$ and variance $s_i (i = 1,2,3,..., n)$

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

3. Classify 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: $(\overline{x_1}, \overline{x_2}, ..., \overline{x_i} \pm s_i, ..., \overline{x_n}), i = 1,2,..., n$

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 2 shows them:

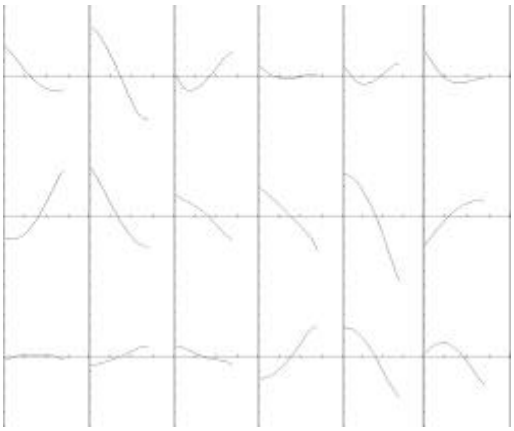

Figure 2: F0 patterns after clustering analysis.

4.DATA MINING

4.1Feature Selection

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. For the F0 pattern classification, the original prosody pattern is divided into three parts: zero-mean F0 pattern, the duration, and the mean as our output. In addition, the linguistic features including tone combination, word length, part-of-speech (POS), syllable position in word, word position in phrase, and the concept class which can be obtained from the text-parsing on a labeled word dictionary.

The original pitch from sentences is discrete with extracted classic F0 models, and at the same time the original length and mean should be kept for future learning. Before training, the Rough set is proposed to find the minimum attribute set. The rough set theory [11] 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 [12]. 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.

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 or decision tree training and testing.

4.2 Learning F0 Model and Length of F0 Based on Decision Tree

A decision tree can be used to classify a case by starting at the root of the tree and moving through it until a leaf is encountered[13]. At each nonleaf decision node, the case's outcome for the test at the node is determined and attention shifts to the root of the subtree corresponding to this outcome. When this process finally leads to a leaf, the class of the case is predicted to be that recorded at the leaf.

Two decision trees are constructed for predicting the F0 model and length of F0 based on C4.5[14]. The condition attributes and decision attributes for F0 model learning are showed in Table1, The condition attributes and decision attributes for Length of F0 learning are showed in Table2. Some rules are also showed.

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

Table1. 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)
	Consonant and tone (pycon, tone)
	pretone, posttone
Decision	Length of F0(discreted)

Table2. Attributes of Decision tree for length of F0

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

Some rules for length of F0 prediction:

wordno=1 and type=14 and pycon=2 and tone=2 -> class 7
type=14 and pycon=2 and pretone=3-> class 7
wordno=2 and len=2 and type=14 and pycon=2
and pretone=5 -> class 5
wordno=1 and type=22 and pycon=2 and tone=2
and posttone=2 -> class 6

4.3 Learning the Mean of F0 Based on ANN

There are many kinds of neural networks, which can be used for learning. Backpropagation is a neural network algorithm for classification, which employs a method of gradient descent. It searches for a set weights which can model the data so as to minimize the mean squared distance between the network's class prediction and the actual class label of data samples. Rules may be extracted from trained neural networks in order to help improve the interpretability of the learned network.

We intend to learn the mapping between the linguistic features and the F0 mean value. Since backpropagation network [15] has implicit input layer and output layer, and it can also give very good result, thus it is chosen to be trained in our system.

For the network learning the mean of F0, 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 3.

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

Table 3: The Definition of Input Layer

After training, the NN and the decision tree can be used to predict and generate the fundamental frequency. F0 model can be modified in accordance with F0 duration and F0 mean. Some experiment result shows in figure 3.

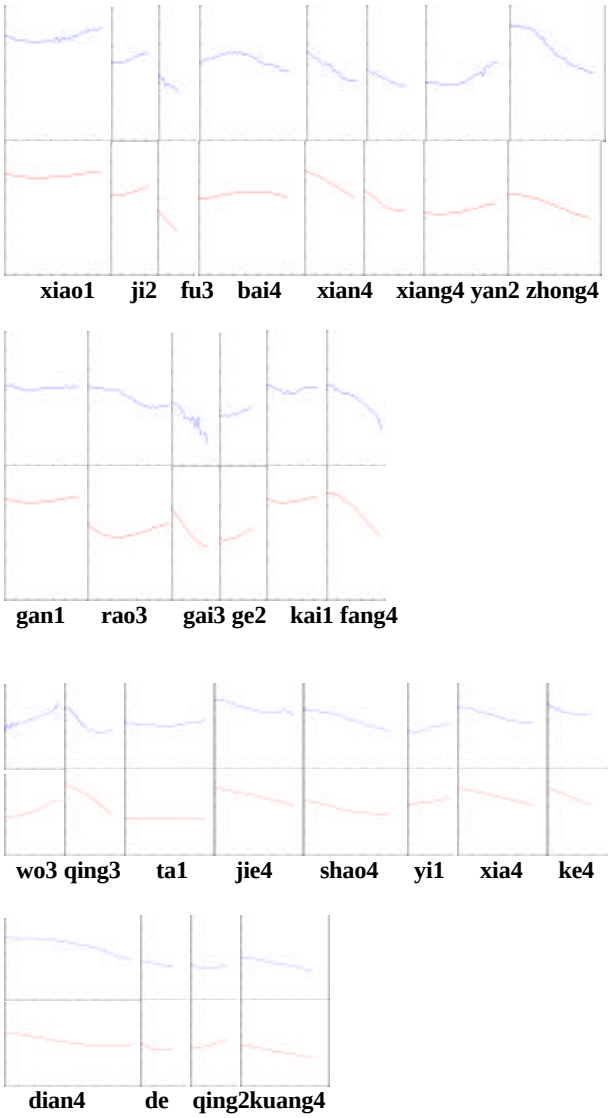

Figure 3 above is original pitch，below is the synthesis one

For most sentences, our experimental results showed that synthesized prosodic features quite resembled their original counterparts.

5.CONCLUSION

In this paper, a Multi-strategy Data Mining framework for extracting prosodic patterns from actual speech database has been proposed. In data preprocessing, some typical prosody models are found by clustering analysis, and these clusters together with some linguistic features obtained by text parsing are used to acquire training data, then Rough Set is employed for feature selection. ANN and Decision tree are trained respectively. 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 by the Multi-strategy mining framework. The experimental results showed that most synthesized sequences match very well with their original counterparts.

6.REFERENCES

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

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Usama M.Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, Ramasamy Uthurusamy. EDITORS, Adavance In Knowledge Dicovery And Data Mining. AAAI/MIT Press, 1996

[9] George H.John, Enhancements to the Data Mining Process, Ph.D thesis of Stanford University, 1997.

[10] Xingjun Yang, Huisheng Chi, Digital Voice Signal Process, Beijing: Publish House of Electronic Industry, 1995.

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

[12] Seoho Kim, Seokkyung Chung, and Mansuk Song, "Rule Acquisition for Nominal Coordinate Structure Based on Rough Sets", *Proceedings of Knowledge Discovery and Data Mining Workshop*, 5th Pacific Rim International Conference on Artificial Intelligence, pp.8-19, 1998.

[13] Chen Wenwei, "Intelligence Decision Technology", *Publishing House of Electronics Industry press*, pp.171-175, 1998.

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

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