Research on predicting prosodic parameters for Chinese synthesis by data mining approach *

WANG Wei CAI Lianhong

(*Department of Computer Science and Technology, Tsinghua University* Beijing 100084)

Received Oct. 25, 2001

Revised Jun. 24, 2002

Abstract Prosodic control is an important part of speech synthesis system. Prosodic parameters choice right or wrong influences the quality of synthetic speech directly. At present, text to speech system has less effective describe to reflect data relationships in the corpus. A new research approach - data mining technology to discover those relationships by association rules modeling is presented. And a new algorithm for generating association rules of prosodic parameters including pitch parameters and duration parameters from corpus is developed. The output rules improve the correctness of syllable choice in text to speech system.

PACS numbers: 43.45

1 Introduction

With the development of the technology in speech processing and computer science, the mandarin text to speech system has been made an important progress, and has been used in various places successfully. But, the results of it is still far away from the high naturalness compared with humans, being lack of the characteristic and rules knowledge of prosodic processing. In the past years, linguist summarized prosodic rules by their phonetic knowledge and engineers induced prosodic rules by statistical method. But, the phenomena of speech are complexity, changeable and humans perception are sensitive, so numerate all rules method is incredible. We need to find a new approach for handling speech prosody. For instance, using data mining approach to mine speech data and discover hidden new prosodic knowledge in the corpus.

Data mining technology is an approach, which is a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. It combines statistical technology and computable technology to discover useful pattern in large database and generate instruction rules. These rules describe data attributes and object sets effectively. Then generate abstract of data sets for decision system. Association rule mining is an important model in data mining technology. Association rules model can reflect relationship of data items in database. Thus, mining Chinese prosodic association rules has important meaning.

* This work was supported by the 863 National High Technology Project and the National Natural Science Foundation of China (No. 60275014).

Compare with prior research of prosodic parameters in text to speech system, this paper has advantage of three aspects: Firstly, we present a new research approach - data mining technology, this method is entirely based on data driven for studying prosodic parameters. Secondly, data mining technology can discover new knowledge in the corpus while past traditional statistical technology can not. Thirdly, data mining technology is suitable to computing large data sets. This method overcome the weakness of rules is given by linguist experience and convert speech data process from subjective to objective.

It is difficult to choose a suitable syllable pronunciation in text to speech system. Under different speech environment information, the same syllable's prosodic parameters represent different value in different sentences in the corpus. Each pronunciation has its own speech environment. The suitable syllable pronunciation is determined by its prosodic parameter values under certain environment. As lacking of efficient description between the previous syllable and the next syllable relationship, it is difficult to predict the next syllable prosodic parameter values even if the current syllable has been set. But the predicting next syllable suitable pronunciation is very important in speech synthesis system. In this paper, we present a new method - data mining approach to mine prosodic parameters in real corpus and generate direction rules.

2 Prosodic parameters

Speech is a physics characteristic sound wave, which produced by a person's voiced organ emit. Each pronunciation has own certain characteristic, which includes timbre, tone, amplitude and duration. Timbre is a basic characteristic feature to differentiate one voice from the other voices. Tone refers to high and lower of a sound. It is also called pitch in Chinese phonetics. The strongness and weakness of a sound is called amplitude. It is determined by wave vibration amplitude. The length of a sound is called duration. It reflects the persistence time of a sound.

Speech prosodic parameters are also called supra segmental feature. In general, it refers to syllable's pitch, duration and amplitude parameters. The Chinese is a tone language, which pitch value is the most important prosodic characteristic. The pitch of tone reflects basic vibration frequency of vocal cords. Different tone has different pitch value. There are all kinds of pitch value in the sensitive machine. Even if the same person pronunciation the same tone word under different speech environment, it shows different frequency. For example: syllable "身" pronunciation has four different frequency: in phrase "身尸" the frequency is 170 Hz, in phrase "身躯" the frequency is 150 Hz, in phrase "起身" the frequency is 160 Hz, in phrase "动身" the frequency is 155 Hz. The person's ear can not recognize the detail of pitch. It often uses a relative physics value to display tone's pitch in linguistic. That is, it can be represented by relative average values. But tone's pitch is not a single frequency value. As tone is continuance frequency values, the frequency has changed from beginning point to ending point during the length of a sound. Through clustering some pitch data experiments, the result shows that cluster pitch vector directly or cluster remove pitch vector, which minus pitch average value, have the same cluster center. So we conclude that use pitch average value as description of prosodic parameters is effective.

The tone is a continuous length wave. Each tone has its own duration. The curve of pitch is beginning with sonorant and ending with sonorant in speech spectrum graph. Linguist mentions that the beginning and the ending curve of a pitch are not useful information in speech perception. The middle curve is becoming useful information, which called efficient duration. Thus, pitch average parameters value and duration parameters are important attributions of a syllable.

In this paper, we use pitch average value parameters as mining object. Through mining the same syllable under different speech environment, we give association rules model to describe relationship between the previous syllable and the next syllable.

3 The algorithm of association rules

The following is a formal statement of the association rules problem: Let $I=\{i_1, i_2, \cdots, i_m\}$ be a set of m literal, called items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$ associated with each transaction is a unique identifier, called its TID. We say that a transaction T contains X, a set of some items in I, if $X \subset T$. An association rule is an implication of the form $X \Rightarrow Y$, where $X \subset T$, $Y \subset T$, and $X \cap Y = \Phi$. The rule $X \Rightarrow Y$ holds in the transaction set D with confidence c if $c\%$ of transaction in D that contain X also contain Y. The rule $X \Rightarrow Y$ has supports in the transaction set D if $s\%$ of transaction in D contain $X \cup Y$. Given a set of transaction D, the problem of mining association rules is to generate all association rules that have support and confidence greater than the use-specified minimum support (called minsup) and minimum confidence (called minconf) respectively. Support controls the minimum number of data cases that a rule must cover and confidence controls the predictive strength of the rule.

The problem of discovering all association rules can be decomposed into two sub-problems:

(1) Find all sets of items that have transaction support above minimum support. The support for an itemset is the number of transactions that contain the itemset. Itemsets with minimum support are called large itemsets, and all other small itemsets.

(2) Use the large itemsets to generate the desired rules. Here is a straightforward algorithm for this task.

Apriori algorithm is a core association algorithm, which has been presented in 1994 by Agrawal etc. Apriori algorithm is a width advantage algorithm. The algorithm for discovering large itemsets makes multiple passing over the database D. Each pass assumption all items that has the same length in the database. In the first pass, the algorithm simple counts individual items and determine which of them are the frequent 1-itemset(with 1 item). In each subsequent pass, we start with a seed set of itemsets found to be large in the previous pass. We use this seed set for generating new potentially large itemsets, called candidate itemsets, and count the actual support for these candidate itemsets during the pass over the data. At the end of the pass, we determine which of the candidate itemsets are actually large, and they become the seed for the next pass. This process continues until no new itemsets are found. Apriori algorithm process objects are no time sequence relationship data, but speech data is a time sequence data.

That is the sequence of speech data influenced the result of data mining directly. We need to improve basic Apriori algorithm. In this paper, we present a new association algorithm (named SPApriori), which is suitable for discovering knowledge of speech data.

SPApriori algorithm description as follows:

(1) C_1 = count(c, D); //count the number of each prosodic parameters occurrences in corpus,

(2) $L_1 = \{c \in C_1 | \text{c.count} \geq \text{minsupport}\}$; //generate larger than minimum support prosodic parameters data,

(3) for $(k = 2; L_{k-1} = \phi; k++)$ do //generate two words or multiple words large itemsets,

(4) for $(k = 0; k < C_1.\text{size}; k++)$ do //sort the generation large itemsets,

(5) $f = C_1[k]$,

(6) if f.count $\geq$ minsupport(f) then,

(7) insert $\{f, f\}$ into C_2,

(8) for $(j = k + 1; j < C_1.\text{size}; j++)$ do,

(9) $g = C_1[j]$,

(10) if g.count $\geq$ minsupport(f) then,

(11) insert $\{f, g\}$ $\{g, f\}$into C_2,

(12) end,

(13) end,

(14) end,

(15) end,

(16) for all transactions $\in D$,

(17) C_t = count_support(C_k, t);// candidate items,

(18) for all transaction $c \in C_t$,

(19) c.count++,

(20) end,

(21) $L_K = \{c \in C_k | \text{c.count} \geq \text{minsupport}\}$,

(22) end,

(23) end,

(24) resultset = resultset $\cup L_k$//generate need rules.

In fact, the core problem is mining all sets of items that have transaction support above minimum support in passing database. If we discover the large itemsets, correspondence association rules are formed easily.

The basic thought of seeking large itemsets is multiple possessing in database. It consists of two phases. First, count the number of each speech data occurrence in the corpus and find that larger than minimum support prosodic parameter data to form 1-dimension large itemsets. Next, This process continues until no new itemsets are found. Loop process is based on $(k-1)^{\text{th}}$ pass generate all frequent $(k-1)$-itemsets to form k-dimension candidate itemsets. Then compare the support of each candidate itemset and minimum support and determine k-dimension maximum itemsets. Here, k-itemset represents k-dimension itemsets. L_k represents

k-itemset larger than what we setting threshold item. C_k represent candidate itemsets. The algorithm terminates when L_k becomes empty.

It is impossible to enumerate all possible association rules since the number of data are too large. Meanwhile the speech data is a time sequence data. The different location has influence of relationship between the current syllable and the next syllable. In this paper we discuss the adjacent relationship, so we set the length is two. The process divides into two steps. Step1 is seeking for current maximum sequence. Step 2 is forming maximum sequence in candidate sequences until these candidate sequences can construct an aggregation. After pass all candidate sequences in database to delete the no-maximum sequences, then form requirement association rules.

4 Experiment research

4.1 Experiment data

Object testing data sets are 2259 Putonghua sentences in our laboratory speech corpus. Experiment corpus is a female sounds database. Sample rate is 16 kHz. The accurate quantization is 16 bits. This corpus includes nearly 20000 syllables. It covers 417 tone syllables and cooperation of multiple sound characteristics. We use speech analysis software SPEECH to tag the boundary and pitch of syllable. Additional, SPEECH software can display the wave of speech data, speech spectrum graph and test any choice sentence in corpus. It is correct syllable database after persons remove some mistakes.

Part of testing data sets formats are shown in Table 1.

Table 1 Part of speech corpus

Serial number 1	—	—	奉	陪	吃	喝				
Pitch average value	291	300	308	161	249	264				
Serial number 2	1	9	9	2	年	体	育	建	筑	奖
Pitch average value	334	263	178	249	172	171	232	329	175	108

After synthesis system word analysis, the sentence format is becoming as follows: —— — 奉陪 — 吃喝, 一九九二年 — 体育 — 建筑奖. Here is only show an example to demonstrate data format in real corpus. We change this data format to what data mining processing needs.

4.2 Pitch data quantization research

Pitch average value data change range has very large in real speech corpus. It is caused by record person whose pronunciation use different speed and tone. So we need to discrete data in different region. Here we use two ways to discrete process: uniformity discrete and normality discrete. Uniformity quantize to 50 region results are shown in Table 2.

Table 2 counts the number of pitch average value data in different region. The data distribution in corpus has some regularity from Table 2. When data locate in region from 9 to 42, the numbers of occurrence are more than other regions. When data locate in region

from 1 to 8 and form 43 to 48, the numbers of occurrence are less. We conclude that the data distribution is not obeying uniformity distribution. The uniformity quantization can not reflect data distribution regularity effectively. We use normality distribution to quantize speech data. Normal distribution result is shown in Table 3.

Table 2 Uniformity quantization and large itemsets 1

Pitch data	Region	Count
84-90	0	/
91-97	1	9
98-104	2	12
105-111	3	25
112-117	4	33
118-124	5	81
125-131	6	115
132-138	7	146
139-145	8	205
146-151	9	220
152-158	10	288
...	...	...
322-328	35	501
329-334	36	458
335-341	37	463
342-348	38	477
349-355	39	386
356-361	40	289
362-368	41	263
369-375	42	200
376-382	43	137
383-389	44	91
390-395	45	43
396-402	46	52
403-409	47	27
410-416	48	9
417-422	49	/
423	50	/

Note: minimum support is 0.5%.

Table 3 Normality quantization and large itemset 1

Pitch data	Region	Count
84-135	0	354
136-148	1	362
149-157	2	371
158-164	3	371
165-170	4	376
171-175	5	298
176-181	6	385
182-187	7	418
188-192	8	361
193-197	9	367
198-202	10	372
...	...	...
318-322	40	384
323-327	41	338
328-332	42	408
333-337	43	367
338-342	44	321
343-348	45	410
349-355	46	386
356-363	47	371
364-374	48	351
375-422	49	388
423	50	354

Note: minimum support is 0.5%.

From Table 3, we can conclude that the number of data in each region is nearly equality. In data sparseness region, partition is becoming large. In data denseness region, partition is becoming small. This means normal distribution is obeying real data distribution.

5 Rules discovery

5.1 Generation rules

First, the minimum support value is given. Minimum support value is specified by user requirement. Then we calculate data occurrence frequency in different region and larger than or equal to minimum support prosodic parameter data as candidate itemsets. For example,

in region 48 and 12, based on SPAprioir algorithm, we can obtain three rules 48→12, 48→20, 48→21.These rules supports are satisfied with threshold support value. The number of support is determined by the value.

48→12, support= 15...

The rule 48→12 meaning is that if current syllable's pitch average value is distribution in region 48, then the next syllable's average value maybe distribution in region 12. This rule has support is 15. The rule support refers to satisfy data occurrence frequency in real corpus.

Table 4 shows the current syllable and the next syllable pitch average value relationship about distribution in region 48 and 49 some rules.

Table 4 the current syllable and the next syllable distribution in region 48 and 49 some rules

Current syllable	The next syllable	Rule support
48	12	15
48	20	10
48	21	12
49	6	10
49	18	10
49	13	10
49	28	10
49	20	11
49	30	12
49	27	16
49	29	10
49	24	11
49	32	10
49	12	11

Table 4 lists some rules about if current syllable pitch value is distribution in region 48 or 49, the next syllable possible distribution region in corpus. The two rules 48→12 and 49→27 have maximum support in generating rules.

These rules mean that if we know the current syllable prosodic parameters for example pitch average value characteristic, then we can predict the next syllable prosodic parameters effectively. These rules will help select the best suitable pronunciation in syllable candidate sets. The next syllable prosodic feature is influenced by different speech environment or sentence stress in sentence to display different prosodic parameter values. We can search for the suitable next syllable by direction association rules. Rules application example gives the pitch characteristics possibility in candidate set in corpus. These rules describe different syllable prosodic parameter values relationship in corpus, and they will help synthesis system to choose the suitable next syllable.

5.2 Rules analysis

When current syllable quantization is distribution in region 48, which correspondence pitch average value is from 364 to 374.All syllables in this region occurrence numbers are 351. The

value is larger than our minimum support cover with threshold value. The minimum support value is equal to all syllables numbers multiple 0.5. In region 48, the pitch average value of syllable "物" is 364. The number of "物" occurrence is 33 in the corpus. The occurrence frequency of the next syllable is shown in Table 5.

Table 5 The occurrence frequency of the next syllable

The next syllable	Pitch data	Count	"质"count number
12	207-211	388	50
20	239-242	384	31
21	243-246	382	36

Table 5 shows that if the current syllable "物" is distribution in region 48, the next syllable "质" occurrence possibility includes 50, 31 and 36. We choose the maximum occurrence rules and put it back in corpus to verify. If this rule exists in corpus, it means this rule is given the best choice of the next syllable prosodic parameter. The other rules can be analyzied in this way. The association rules reflect different speech data item relationship in sentence.

5.3 Rules application example

We use some examples to demonstrate this method in our real text to speech system. For example, phrase "优质苦咖" and "坚持物质文明" after text analysis, the phrase data format convert to word segment format — 优质 — 苦咖 — and — 坚持 — 物质 — 文明 —. We use the word segment result as the input of SPAprioir association algorithm. We adapt normality distribution to pitch average value, some rules shows in Table 4. The syllable "物" pitch average value is distribution in region 48 by normality quantization. The rule 48→12 is the maximum support rule. Search in our speech corpus, the syllable "质"is distribution in the region 12. This rule shows if previous syllable "物" distribution in region 48 then we should choose "质" which distribution in region 12 not in region 32. The other rule 49→27 is another maximum support in Table 4. If syllable "优" in region 49 can not find correspondence the next syllable "质", which in region 27 in corpus. We need decrease the minimum support until obtain another rules. Then we find another rule 49→32 which exists in corpus. The syllable "质" representation different pitch average value is caused by different speech environment.

The mining association rules reflect the relationship between adjacent syllable pitch average values. These rules will help to choose suitable syllable effectively and improve the accuracy of selection pronunciation.

Fig. 1 shows the wave of two phrases pitch average value by initial record person pronunciation. The word "物质" divide into "物" and "质" which pitch average value is 364, 207 respectively. After quantization, the distribution region is 48, 12 respectively. The word "优质" divide into "物" and "质" which pitch average value is 375, 288 respectively. After quantization, the region is 49, 32 respectively. Fig. 1 illustrates if the previous syllable is "物", the next syllable selection is "质", we choice syllable pronunciation is distribution in region 12 not in region 32.

Finally, we need point out that Table 4 experiment results are limited by research corpus. It can not cover all kinds of possibility of certain data in all kinds of corpus. While, the certain result under certain corpus shows this method is effectively. We believe that if corpus covers syllables pronunciation under different speech environment as many as possible, the mining result will be better.

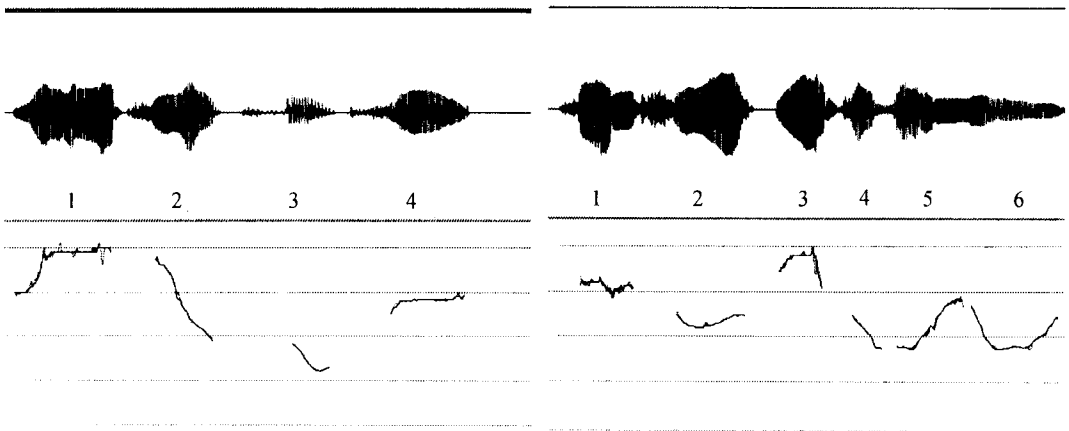

Fig. 1 The pitch average value wave of phrases"优质苦咖"and"坚持物质文明"

6 Conclusion

In text to speech system, the relationship between syllables has influence about the text to speech output quality directly. In this paper, we present a new method - data mining approach apply into Putonghua prosodic feature includes pitch average parameters and duration parameters for knowledge discovery. We present a new association algorithm for discovering new knowledge in the corpus. This method can discover new knowledge while statistical method can not. The rules can direct system choice the best suitable pronunciation in candidate syllables sets. But the minimum support influences the result directly. How to choose the best minimum support is our future research task.

References

[1] GUO Jinfu. Hanyu shengdiao yudiao chanyao yu tansuo. Beijing Language and Culture Press, 1993(in chinese)

[2] YAO Tianren. Digital speech processing. Huazhong Institute of Technology Press, 1992(in chinese)

[3] Rakesh Agrawal, Ramakrishnan Srikant. Fast algorithm for mining association rules. In: Proceedings of the 20th VLDB Conference, Santigo, Chile, 1994

[4] Ramakrishnan Srikant, Rakesh Agrawal. Mining generalized association rules. In: Proceedings of the 21st VLDB Conference, Zurich, Switzerland, 1995

[5] Rakesh Agrawal, Tomasz Imielinski, Arun Swami. Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD Conference, Washington DC, USA, 1993