ANNOTATION OF CHINESE PROSODIC LEVEL BASED ON PROBABILISTIC MODEL

Rui CAI

Department of Computer Science and Technology,
Tsinghua University,100084, Beijing
cairui01@mails.tsinghua.edu.cn

Zhi-Yong WU

Department of Computer Science and Technology,
Tsinghua University,100084, Beijing
wuzy99@mails.tsinghua.edu.cn

Lian-Hong CAI

Department of Computer Science and Technology,
Tsinghua University, 100084,Beijing
clh-dcs@tsinghua.edu.cn

ABSTRACT

In this paper, a probability based method is proposed for the annotation of Chinese prosodic levels. We investigated the acoustic correlates (F0 pitches, durations and pauses) of the prosodic boundaries and tried to annotate the levels of the boundaries using Gaussian model and Bayesian decision. The method allows efficient and automatic labeling for the large scale speech corpus and can be used to instruct the unit selection in TTS system.

1. INTRODUCTION

In the past few decades, the Chinese TTS technique has greatly developed. Through using smooth-operation and rules-control, the problem of naturalness has been improved by some degrees. Nevertheless, the characteristic of Chinese has been imitated insufficiently and far from the actual speech. Chinese prosodic levels play an important role during the communication, which have been actively researched and formally addressed by researchers. On the one hand, the corpus should have been labeled with prosodic information before being used in synthesis systems and on the other hand, results of analysis can be used to instruct the unit selection.

The early works focused on analyzing from text to predict prosodic boundaries [3]. To date, numerous attempts have been made to achieve the prosodic characteristic through analyzing perception of speech. Wei-Xiang HU et. al. [1] proposes a D.T. (Decision Tree) method which can automatically determine Chinese prosodic level by employing 26 features. In our experiments, we extract less number of acoustic features and adopt probability based method.

Four types of prosodic levels were concerned in our experiment, which were represented by random variable **T** in this paper: (1) syllable boundary **A**, (2) word boundary **B**, (3) prosodic chunk boundary **C**, and (4) intonation phrase boundary **D**. Given a median scale female speech corpus with news-report style, which contains 3000 sentences and about 20 syllables (characters) for each sentence, the prosodic boundaries of the corpus are labeled manually, results from different testers have been examined for certain consistency. From the above labeling data, we studied several acoustic characteristics of prosodic boundaries, and have found that the prosodic levels of boundaries have great concern to the fundamental frequency (F0 pitch) of the syllables around the boundary.

Three types of features were computed and analyzed for each boundary: **(I)** the duration of the F0 contour of the pre-boundary syllable, **(II)** the F0 pitch reset for post-boundary and pre-boundary, and **(III)** the duration of the silence of the pitch contours between post-boundary and pre-boundary, which respectively represented by random variable **X**, **Y**, **Z** in this paper. For certain prosodic boundary, we found that the conditional probability distribution of each feature submits to the Gaussian distribution, the conditional jointly probability distribution of every two features consists with 2-D Gaussian distribution, and the jointly distribution of all three features with 3-D Gaussian distribution well. The expectation and standard deviation of each Gaussian distribution can be estimated by statistical method, which will be used in the Bayesian decision method.

The feasibility of the proposed probability based method was demonstrated in our experiments. The Chinese prosody boundary can be well described using the three features above. The annotation of prosodic level using Gaussian model and Bayesian decision has good performance.

2. EXPERIMENTAL FRAMEWORK

The experiment corpus used in this work is a median scale female speech corpus with news-report style, which contains 3000 sentences and about 20 syllables (characters) for each

sentence. The prosodic boundaries of the corpus are labeled by 10 members of our lab without text information. We represent different boundaries with score 0~3, and take the round of average score of all testers as the level of the boundary.

The data is divided into two parts: training data and test data. The training data is consisting of 2000 sentences and the remainders are used for test.

Table 1 gives the essential acoustic parameters of this corpus.

Tone	One	Two	Three	Four	Avg.
F0 (Hz)	295	236	200	236	251
Dur. (ms)	256	255	247	234	247

Table 1 Corpus's essential acoustic parameters

3. FEATURE EXTRACTION

3.1 Duration of Pre-Boundary Syllable

The pre-boundary lengthening was considered as an important acoustic parameter of prosodic boundary, especially for the weak boundaries. In our experiment, we use F0 contour duration instead of syllable duration (Fig. 1) for two reasons. One is because the detection of begin/end point of F0 contours is more accurate using automatic label. The other is because the syllable lengthening is vowel lengthening and F0 contours can describe vowel's length accurately.

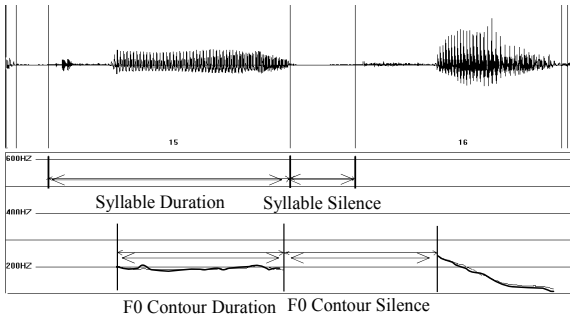

Figure 1 Difference of Syllable and F0 Contour's Duration / Silence

3.2 Duration of Silence

Insertion of silence is another important feature of the prosodic boundaries, especially for the strong boundaries. The higher the level of prosodic boundary processes, the longer the duration of silence persists. We use F0 contour's silence instead of syllable silence (Fig. 1) for the same reason shown in section 3.1.

3.3 F0 Pitch Reset

Discontinuation of intonation's bottom line, which is composed of low points of the pitch, is a crucial feature to all kinds of boundaries. We define pitch reset value as the difference of low points of post-boundary and pre-boundary pitch (Fig. 2 Left).

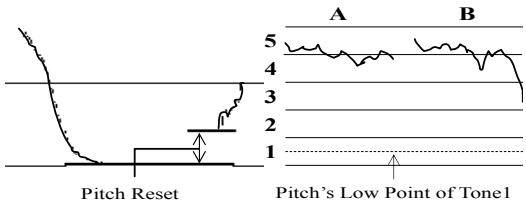

Figure 2 Definition of pitch reset (left) and pitch's low point of tone 1 (right)

There is a problem on how to deal with tone 1's pitch low points, which can not represent the lower limit of tone 1's range. The average value of pitch's low points of tone 1 is much higher than the other three tones (Table 2). It's illogical to use this value directly.

Low Point	Tone 1	Tone 2	Tone 3	Tone 4
Avg. (Hz)	254	196	173	198
σ (Hz)	27.5	20.1	16.4	22.3

Table 2 Distribution of pitch's low points of different tones

We provide a solution here as shown in Fig 2 (Right). Given a syllable's pitch contour of tone 1, we calculate its stand deviation σ firstly. If σ is lower than a predefined threshold σ_0, the syllable's range will be considered between 4 and 5 in logarithmic coordinate, otherwise the range will be considered between 3 and 5. After that, we can achieve the value of middle position of region 1 (the dashed in Fig. 2). Finally, this value was converted back to original frequency coordinate and we take it as the pitch's low point.

4. GAUSSIAN MODEL AND BAYESIAN DECISION

4.1 Feature Distribution and Gaussian Model

The statistical results of the training set (2000 sentences) are shown in Table 3.

Boundary Level		A	B	C	D
Total Number		15808	7884	5308	702
Duration of Pre-boundary	Avg.(ms)	216	235	297	315
	σ (ms)	47	51	54	52
Pitch Reset	Avg.(Hz)	-12	4	19	32
	σ (Hz)	33	36	35	36
Silence	Avg.(ms)	87	115	152	369
	σ (ms)	66	58	55	97

Table 3 Statistical results of feature distribution

It's clear that the higher the prosodic level is, the larger these features' values are. It's also clear that the average of duration of the pre-boundary syllables increase little between level C and D, while the average of silence increase notably. These data prove those theories referred to in section 3.

Given a certain boundary level, we can get the illustrations of each feature's distributions and every two features' joint distributions by MATLAB (Fig. 3), which can be treated as Gaussian distribution.

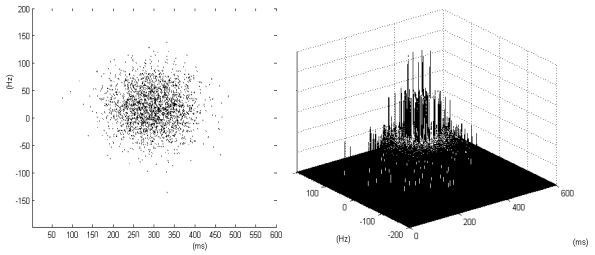

Figure 3 2D and 3D illustrations of Pitch Reset and Pre-Boundary Syllable Duration's Joint Distribution

It's hard to give the illustration of 3-D joint distribution of all the three features. We assume it can be treated as 3-D Gaussian Model, which will be proved to be reasonable by the experiment's result.

4.2 Bayesian Decision

The process of automatic annotation of Chinese prosodic level in fact is a process of classification. Given some obvious features and judge which class the sample is belongs to by some rules or models. Statistic theory is one of the essential theories of pattern recognition which is often used to solve problems with large scale training data and Bayesian decision theory is the most essential theory of statistical pattern recognition.

4.2.1 Prior Probability

In the condition of large scale corpus, we take it for granted that the occurrence probabilities of each level can be treated as a constant for a certain speaker in a certain style (such as news report style).

In our experiment, we use random variable **T** to represent prosodic boundary level. The possible value of **T** can be **A~D**. The prior probabilities of all kinds of prosodic boundaries in our training set are listed in table 4.

t	A	B	C	D
Pr(T=t)	53.22%	26.54%	17.87%	2.37%

Table 4 Prior probability of different boundary

4.2.2 Decision Method

Random variables **X**, **Y** and **Z** were used to represent pre-boundary syllable's duration, pitch reset and silence respectively. Given a certain syllable boundary, we can get the value of each random variables above, which was represented by **x**, **y**, and **z**. Then **t** corresponds to the maximum value of the conditional probability

$$\Pr(T = t \mid X = x, Y = y, Z = z) \quad t \in [\mathrm{A},\ \mathrm{B},\ \mathrm{C},\ \mathrm{D}] \qquad (1)$$

represents the predict result.

(1) can be transformed by Bayes' formula into the following form:

$$\Pr(T=t \mid X=x, Y=y, Z=z) = \frac{\Pr(X=x, Y=y, Z=z \mid T=t)\Pr(T=t)}{\Pr(X=x, Y=y, Z=z)} \qquad (2)$$

The denominator in (2) can be transformed by total probability formula:

$$\Pr(\mathrm{X = x, Y = y, Z = z}) = \sum_{t} \Pr(\mathrm{X = x, Y = y, Z = z \mid T = t})\Pr(\mathrm{T = t}) \quad (3)$$

As we assumed in section 4.1, the joint distribution of the three features for a certain kind of boundary can be treated as 3-D Gaussian Model. It's easy to represent 1-D G.M. by its mean **μ** and standard deviation **σ**. The 3-D G.M. can be represented by $\vec{\mu}$ and **Σ** similarly. The difference is that $\vec{\mu}$ is a vector and **Σ** is a symmetric matrix. And the ***p.d.f.*** (probability density function) of 3-D G.M. can be represented as follows:

$$f_{X,Y,Z|T=t} = \frac{\exp\{-\frac{1}{2}(\vec{x} - \vec{\mu}_t)^T \Sigma_t^{-1}(\vec{x} - \vec{\mu}_t)\}}{(2\pi)^{3/2} |\Sigma_t|^{1/2}} \quad (4)$$

where

$$\vec{x} = (x, y, z)^T, \vec{\mu}_t = (\mu_{X_t}, \mu_{Y_t}, \mu_{Z_t})^T, |\Sigma| = \det\Sigma$$

The conditional probability can be represented by ***p.d.f.***:

$$\Pr(\mathrm{X = x, Y = y, Z = z \mid T = t}) = f_{X,Y,Z|T=t}(X, Y, Z \mid T) * \Delta V \quad (5)$$

where ΔV represents the unit in feature space. Finally, equation (2) was transformed by (3) (5) to:

$$P(T = t_0 \mid X = x, Y = y, Z = z) = \frac{f_{X,Y,Z|T=t_0}(x, y, z)P(T = t_0)}{\sum_t f_{X,Y,Z|T=t}(x, y, z)P(T = t)} \quad (6)$$

We can get $\vec{\mu}$ and **Σ** by statistical method using the training data and statistical results are listed in table 5.

Boundary		A	B	C	D
μ	X	215.543	234.988	297.137	314.592
	Y	-11.819	3.659	19.472	32.311
	Z	86.930	114.528	151.854	369.175
Σ		Σ_A	Σ_B	Σ_C	Σ_D

$$\Sigma_A = \begin{bmatrix} 2286.5 & 15.3 & 121.7 \\ 15.3 & 1145.6 & 45.2 \\ 121.7 & 45.2 & 4388.1 \end{bmatrix} \quad \Sigma_B = \begin{bmatrix} 2679.1 & 5.6 & 0.6 \\ 5.6 & 1362.9 & -13 \\ 0.6 & -13 & 3413.7 \end{bmatrix}$$

$$\Sigma_C = \begin{bmatrix} 2996.7 & 40.6 & -50 \\ 40.6 & 1236.1 & 55.3 \\ -50 & 55.3 & 3110.2 \end{bmatrix} \quad \Sigma_D = \begin{bmatrix} 2716.1 & -119 & -295 \\ -119 & 1331.4 & -331.2 \\ -295 & -331.2 & 9432.1 \end{bmatrix}$$

Table 5 Model's parameters with training set

4.2.3 Probabilistic Error Estimates

According to the probabilistic model, the decision space was divided into four sub-spaces represented by $\mathbf{R_A}$, $\mathbf{R_B}$, $\mathbf{R_C}$ and $\mathbf{R_D}$. The probabilistic accuracy can be defined as follows:

$$P(C) = \sum_{t\in\{A,B,C,D\}} \iiint_{R_t} f_{X,Y,Z|T=t}(x,y,z)P(T=t)dxdydz \quad (7)$$

It's easy to get the estimation of probabilistic error:

$$P(E) = 1 - P(C) \quad (8)$$

In our experiment, the estimation of probabilistic error is about 25.8%. We find most errors are mistake predictions on successive levels, such as boundaries belong to level **B** were predicted to belong to level **A**, which can be accepted in most conditions.

5. RESULT

Given a sentence in the test set:

"市[1]长[2]/关[3]永[4]光[5]|对[6]他[7]说[8]//咱[9]俩[10]/

别[11]让[12]/专[13]家[14]们[15]|上[16]/咱[17]这[18]|领[19]了"

This sentence has been labeled manually as shown above in which token '|' represents word boundary, '/' represents prosodic chunk boundary and '//' represents intonation phrase boundary. Table 6 gives the probabilities for prediction.

No	X	Y	Z	Pr (T=A)	Pr (T=B)	Pr (T=C)	Pr (T=D)	
1	115	−66	55	0.8991	0.1003	0.0006	0.0000	A
2	241	61	105	0.2850	0.4727	0.2171	0.0252	B
3	181	−73	9	0.9291	0.0703	0.0006	0.0000	A
4	242	−43	98	0.7026	0.2605	0.0357	0.0012	A
5	195	54	102	0.4543	0.4647	0.0752	0.0058	B
6	88	−5	135	0.7132	0.2804	0.0064	0.0000	A
7	87	14	141	0.6474	0.3425	0.0101	0.0000	A
8	250	35	392	0.0530	0.0224	0.1236	0.8009	D
9	182	−35	15	0.9026	0.0957	0.0015	0.0002	A
10	237	34	130	0.3819	0.4151	0.1925	0.0105	B
11	157	−8	15	0.8869	0.1114	0.0015	0.0002	A
12	226	58	121	0.3158	0.4769	0.1922	0.0151	B
13	180	−12	57	0.8137	0.1795	0.0064	0.0004	A
14	145	−61	9	0.9369	0.0628	0.0003	0.0000	A
15	197	37	165	0.4273	0.4299	0.1383	0.0045	B
16	276	12	147	0.2885	0.3401	0.3538	0.0176	C
17	203	0	54	0.7663	0.2192	0.0133	0.0012	A
18	143	11	15	0.8630	0.1351	0.0017	0.0002	A
19	191	−18	219	0.6091	0.3097	0.0796	0.0016	A

Table 6 A sample of prediction by probability model

The result of hierarchical boundaries is:

"市长|关永光|对他说//咱俩|别让|专家们|上/咱这领了."

Table 7 gives the accuracy on the test set.

Manually Labeled	**A**	**B**	**C**	**D**
Accuracy	94.23%	76.12%	58.61%	86.47%

Table 7 Accuracy on test set

After analyzing the predict result we find that 1) accuracies of boundaries belongs to level **A** and **D** are satisfied, 2) boundaries in level **B** were used to be classified into level **A** (about 20% of boundaries in level B) and similarly 3) boundaries in level **C** were used to be classified into level **B** (about 38% of boundaries in level C). Actually, it's not easy for human beings to distinguish between them so that thus mistakes are tolerance in most conditions.

6. CONCLUSION

In this paper, we've presented a novel method for annotation of Chinese prosodic levels. The solution employs Bayesian Decision and Gaussian Model.

We concerned four types of prosodic levels named syllable boundary, word boundary, prosodic chunk boundary, and intonation phrase boundary. And we extracted three types of acoustic features for each boundary including the duration of the F0 contour of pre-boundary syllable, the F0 pitch reset, and the duration of the silence of the pitch contours between post-boundary and pre-boundary.

We conducted an experiment that allowed us to show how to estimate the prosodic level using probabilistic method. The feasibility has been demonstrated in our experiments. The Chinese prosody boundary can be well described using the three features above and the prediction results let us satisfied.

7. REFERENCES

[1] Wei-Xiang HU, Bo XU, Tai-Yi HUANG. *Research on detection and recognition on prosodic boundaries of Chinese*. NCMMSC6, 39-42, 2001.

[2] Bei WANG, Yu-Fang YANG, Shi-Nan LV. *Acoustic correlates on prosodic hierarchical boundaries of Chinese.* The Proceeding of 5th National Conference on Modern Phonetics, 161-165, 2001.

[3] Xi-Ping SHEN, Bo XU. *A CART-based hierarchical stochastic model for prosodic phrasing in Chinese*. ISCSLP00, 105-109, Beijing, 2000.

[4] Mao-Chan LIN. *The acoustic manifestation of prosodic phrase boundaries in standard Chinese.* Proceeding of Conference on Phonetics of the Language in China, City University of Hong Kong, 1998.

[5] Marc S. *Prosodic features at discourse boundaries of different strength.* J. acoust. Soc. Am. 101(1), 514-521, 1997.