Prosody Variation: Application to Automatic Prosody Evaluation of Mandarin Speech

Huibin Jia[1], *Jianhua Tao*[1], *Xia Wang*[2]

[1]National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing, China, 10080
[2]Nokia Research Centre, China
[1]{hbjia; jhtao}@nlpr.ia.ac.cn, [2]xia.s.wang@nokia.com

Abstract

Prosody evaluation is an essential part of computer-aided language learning system. In the paper, prosodic variability among inter-speakers is investigated based on a database containing eight repetitions of 200 sentences. For Mandarin of reading style, its variability can be analyzed from rhythm, intonation and tone. Experimental results show that the mean correlation of tone between inter-speakers is 0.70, intonation and rhythm are 0.81. Based on these analyses, the prosodic similarity between the tested and standard utterances is calculated to automatically evaluate prosody quality. The standard utterances were recorded by multiple speakers, so they can cover different prosody patterns for the same utterance. The prosodic similarities are calculated from three aspects: tone, intonation and rhythm. Based on these similarities, the prosody quality can be graded. The method evaluated on the collected database has achieved good performance, and the correlation of human-machine scores is close to that of human-human scores.

1. Introduction

As the Chinese language learning has become more and more popular in the world, computer-aided language learning (CALL) systems in Chinese have become more and more important [3]. Via CALL systems, learners can get real feedbacks of their pronunciation quality online. However, most of CALL systems for Mandarin Speech just focus on the pronunciation evaluation with the integration of automatic speech recognition (ASR) technology [5, 8], e.g., the Mandarin level evaluation system [4] which evaluates native learners' pronunciation with dialect accent. Actually, prosody components are very important for both comprehension and intelligibility. For example, a non-Chinese speaker can be considered as a high-level speaker of Mandarin if he/she can simulate very native and natural Mandarin prosody in continuous speech. Currently, although, several systems are able to detect the syllabic tone alone, there is no system which can automatically evaluate the overall prosody quality (naturalness).

In general, it is very difficult to evaluate prosody quality by directly computing prosodic errors. Because there is no clear definition of "correct" prosody in theory and even a large prosodic deviation from the theoretical standard does not necessarily result in something "wrong". Furthermore, for the same utterance, there are multiple prosody patterns. In [7], the two repetitions (separated by a time span of 6 months) were recorded by a single professional speakers, who was instructed to read these sentences in the same reading styles, and it is observed statistically that the two repetitions have fairly wide variations in prosodic features and even the variations can be up to 50% of the full dynamic range of the speaker. Besides, the complex nature of prosody and its vast interactions with other components make the prosody evaluation very difficult. So far, little work has been carried out. Therefore, new methods must be devised to automatically evaluate prosody quality. In the paper, first of all, prosody variation of inter-speakers is investigated from tone, intonation and rhythm. The experiment is based on four repetitions of 200 sentences. The experimental results show that the average correlation of tone of inter-speakers is 0.70, intonation and rhythm are 0.81. Due to the prosody variation, two metrics for measuring the prosodic similarities between the tested utterance and its standard ones are proposed. Then the two metrics are applied to an automatic prosody evaluation method. In our method, the standard utterances were recorded by multiple speakers, so they can basically cover all of prosody patterns for the same utterance. The prosodic similarities are calculated from three aspects: tone, intonation and rhythm. Finally, experiments on the collected database show that the method has achieved good performance and the correlation of human-machine scores is very close to that of human-human.

The paper is organized as follows: Section 2 describes the collected database. In Section 3, a detailed analysis of prosodic variation is given. Based on these analyses, the similarity between the tested and its standard utterances can be calculated. The experimental results are given in Section 4. The final section concludes with the experimental results.

2. Database

For exploring the prosody variation via automatic prosody evaluation, two databases are constructed. One is the native database which provides the criterion for the tested utterances and it is also used for the analysis of prosody variation. The other, which is evaluated by expert raters at different levels of prosody details, contains all types of prosody errors. These two databases are used to develop and calibrate the prosody scoring algorithm.

In the paper, we only focus on prosody evaluation for Mandarin reading speech and do not involve pronunciation quality evaluation. Recently corpus-based text-to-speech systems have made much progress, and the synthesized speech only has various types of prosody errors and does not have pronunciation errors. Consequently, the synthesized speech is very suitable for the prosody evaluation task.

The synthesized database comes from several text-to-speech systems. It consists of 1200 utterances and each utterance contains about 14 syllables. Besides, the database includes various prosody errors, such as inappropriate tone variation, intonation and rhythm, which are evaluated by expert raters. The native database with standard prosody consists of 6 women's speech and 2 men's speech; the eight repetitions of 200 utterances were recorded by multiple professional speakers, who were instructed to read these sentences in the same reading styles. Both the databases are automatically segmented and pitch-extracted and then are modified by hand. The segmented unit is syllable.

3. Prosody Variation Analysis

Prosodic information usually cannot be associated with a single phone-sized segment, it is referred to as supersegmental information. Speakers use prosody to convey emphasis, intent, attitude, and to provide cues to aid listeners in the understanding of their speech. Prosodic planning at all these levels is realized mainly in three acoustic parameters: pitch, duration and intensity. Generally, there are some differences between prosodic patterns even when different speakers utter the same utterance due to the complicated tone variation, flexible intonation and diverse types of rhythm patterns. In the paper, the analysis and evaluation is only focused on reading speech. The prosody can be decomposed into tone, intonation and rhythm and do not include sentence stress.

3.1. Prosody Modeling

Among these acoustic parameters mentioned above, the pitch is the most important clue. Before the pitch is used for tone and intonation modeling, it is normalized with Equ.1 to reduce the effects of inter-speakers according to the speaker's pitch range.

$$\overline{f_0} = \frac{f_0 - \mu_{f_0}}{\sigma_{f_0}} \qquad (1)$$

Where μ_{f_0} is the mean of $f0$ range and σ_{f_0} is its standard deviation.

3.1.1. *Tone modeling*

The fundamental frequency curve of the entire syllable is divided into several subsections and each subsection is represented by a linear function. The one-order linear function is represented by the slope S_{F0}^{K} and intercept $I_{F0}{}^{K}$. That is, using $f(t) = S_{F0}^{K} t + I_{F0}{}^{K}$ to approximate the F0 curve which belongs to this subsection. As shown in Fig.1, the fundamental frequency curve of "jing1" is divided by 4 equal-length subsections, and each subsection is represented by a one-order linear function. Consequently, the syllabic tone is represented by four pairs of the linear function parameters.

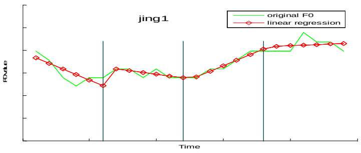

Fig.1 Tone Modeling

3.1.2. *Intonation modeling*

Intonation can be represented by three pitch curves which consist of maximal, average and minimal pitch value of each syllable, respectively [1]. The Fig.2 shows the three pitch curve and the real pitch curve. From the Figure, the three curves can elegantly denote the intonation.

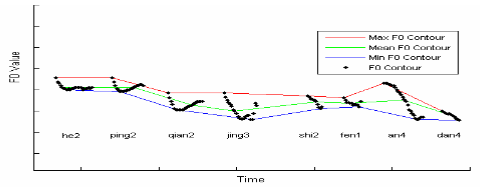

Fig.2 Intonation Model

3.1.3. *Rhythm Modeling*

Rhythm is also one of important prosodic components. The rhythm of one utterance is extracted from the utterance and can be represented by a numeric sequence, such as $\{0,1,2,3...2,1...\}$. In such sequence, 0 stands for no breaks between the two syllables, 1 for prosody word boundary, 2 for prosody phrase boundary and 3 for intonation phrase boundary.

3.2. Variability of Prosodic Components

This section analyzes the prosody variability in Mandarin. In general, speakers have to obey constraints on various levels when speaking. Some constraints, such as the speed of the movement of articulators or the highest/lowest frequency of the vibration of vocal cord are physiological constraints that speakers are not surpass, others, such as linguistic or affective constraints, are human-initiated ones that serve delivering message [7]. These factors results in different prosody patterns of the same text uttered by multiple speakers. For Mandarin, the prosody variation can be analyzed from tone, intonation and rhythm. In the experiment, four repetitions from the standard database are analyzed. The correlation of prosodic component between inter-speakers is calculated. The result is listed as follows:

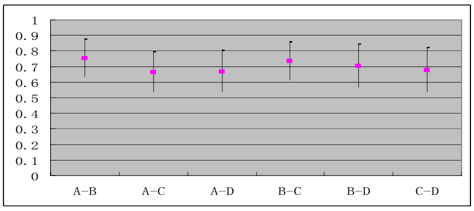

Fig.3 Mean and Standard Deviation of Tone Correlation Between speakers

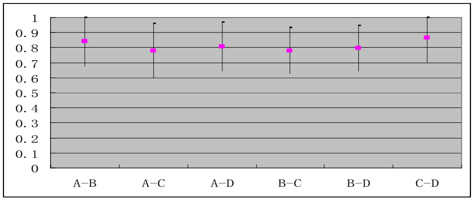

Fig.4 Mean and Standard Deviation of Rhythm Correlation Between Speakers

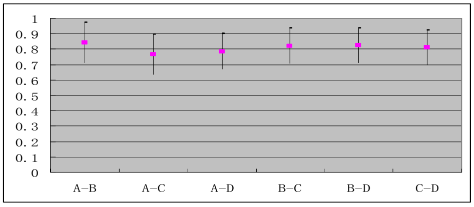

Fig.5 Mean and Standard Deviation of Intonation Correlation Between Speakers

3.2.1. Tone Variation

Each syllable in Mandarin has fixed time-varying pitch pattern, which is lexically given and can not be changed arbitrarily. Stead, in real context, the exact pitch movement of individual syllables is far deviated from their lexical patterns due to the phonological constraints and limitations in the phrase and sentence level.

However, for the tone variation of inter-speakers, these tones have basically the same contour of pitch pattern due to their context and given lexical tone, while the differences mainly exist in detailed pitch curve. The average correlation of four repetitions is 0.70, and the distribution of correlation is shown in Fig3

3.2.2. Intonation Variation

The intonation can be modeled in Section 3.1.2. In Fig 2, the three lines can basically represent the intonation contour. The distribution of intonation correlation between repetitions is shown in Fig5. The average correlation is 0.81.

3.2.3. Rhythmic Variation

Rhythm is one of the most important prosodic components. When different speakers utter the same utterance, maybe they break the sentence into different appropriate rhythmic units. That is to say, there may exist more than one liable ways to break a sentence. Multiple ways of breaking are often acceptable and they sound equally natural to human ears. Specifically, there are multiple breaks allocations in one sentence. Speakers utter it with random combination patterns, namely various rhythm patterns result from randomly choosing break allocation. Besides, the meaning of the sentence is not changed. Such can be shown as follows:

亚洲|金融风暴$把东南亚经济|搞跨了。
亚洲金融风暴|把东南亚经济|搞跨了。
亚洲 金融风暴$把|东南亚经济 搞跨了。
亚洲 金融 风暴|把东南亚 经济|搞跨了。
亚洲 金融风暴$把 东南亚 经济|搞跨了。
亚洲 金融 风暴|把 东南亚 经济|搞 跨了。

Where ‘ ’ stands for the prosodic word boundary, ‘|’ for the prosodic phrase boundary and ‘$’ for the intonation phrase boundary. In the paper, the correlation distribution of four repetitions is shown in Fig 4. The average correlation is 0.81.

3.3. Prosodic Similarities

As mentioned above, there may be multiple appropriate prosody patterns for one utterance, which are equally natural to listeners. When the prosody quality of the tested utterance is evaluated, how to choose the standard utterance as its reference due to the prosodic variation is a very important problem. According to the prosodic variations described above, we can calculate the prosodic similarities (consisting of tone, intonation and rhythm similarities) between the utterance and its references. In the paper, two methods are adopted to compute the prosodic similarities---“Optimal Similarities” and “Weighted Similarities”.

3.3.1. Optimal Similarities

Firstly, the overall prosodic similarity S_j between the utterance and its jth reference is obtained through Equ.2.

$$S_j = \sum_{i=1}^{3} w_i L_{i,j} \qquad (2)$$

Where $L_{i,j}$ denotes the similarity of the i th prosodic component between the utterance and its j th reference and w_i is its weight. $L_{i,j}$ is calculated from the correlation coefficient equation and the setting of w_i is based on experience and $\sum_{i=1}^{3} w_i = 1$.

Secondly, we can choose the maximal S_j through Equ.3. Its corresponding reference is the standard reference whose prosody pattern is closest to that of the tested utterance. The corresponding prosodic similarities become “Optimal Similarities” between the utterance and its references.

$$S_{\max} = \underset{j}{\overset{J}{Max}} S_j \qquad (3)$$

Where J is the number of references.

3.3.2. Weighted Similarities

From Section 3.3.1, the overall prosodic similarity S_j between the utterance and its j th reference is obtained. Based on Equ.4, the weighted similarities of i th prosody component between the utterance and all of its references can be obtained.

$$\overline{L}_i = \sum_{j=1}^{N} w'_j L_{j,i} \qquad (4)$$

$$w'_j = \frac{S_j}{\sum_{i=1}^{N} S_i} \qquad (5)$$

Where $L_{j,i}$ is the similarity of the i th prosody component between the utterance and its j th reference, and w'_i , obtained by Equ.5, is the weight of the corresponding similarity. In addition, N is the number of the reference in the standard database.

4. Experiments and Conclusions

4.1. Human Scoring

Human scoring of prosody quality is a highly subjective task. When multiple experts rate the same utterance, they maybe allocate different scores. The human scores are the reference against which the automatic scoring algorithm should be tested and calibrated. In the paper, six experts are selected from a group of nine candidates as the most self-consistent raters. This panel of six experts rated the overall prosody quality of each of the 1200 utterances on a scale of 1-5, ranging from the categories “very bad” to “excellent”.

We computed the correlation between the scores of a rater and those of the mean of all other raters excluding the current one, which is referred to “open correlation”. The mean of these correlations suggests an upper bound on the level of correlation between human and machine scores.

Table 1 Sentence-level Correlation

Rater ID	1	2	3	4	5	6	Average
Correlation	0.83	0.83	0.70	0.78	0.71	0.70	0.76

4.2. Automatic Prosody Scoring

A block diagram of the automatic scoring method is illustrated in Fig.6. For each tested utterance, its prosodic parameters are firstly extracted and analyzed. Secondly according to these parameters, the prosodic similarities between the utterance and its references are calculated from tone, intonation and rhythm. The references from the standard database have native prosody patterns with the same text. Finally, on the basis of those prosodic similarities, the prosody score of the utterance is predicted by the ranking model. The method is in detail described in [11].

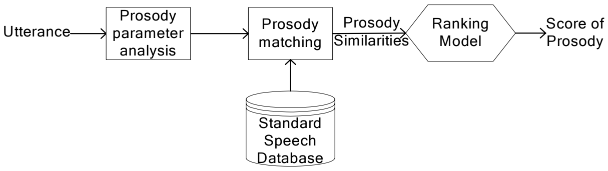

Fig 6 Framework of Scoring System

4.3. Experimental Results

The performance of the scoring system varies with different numbers of references and different distance metrics. The experimental results are shown in Fig.5.

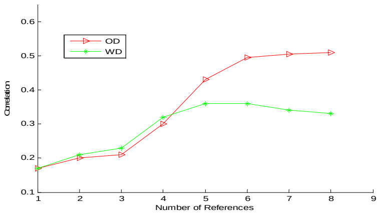

Fig.7 Distribution of Correlations

In the standard database mentioned in Section 2, each of eight different native speakers spoke 200 utterances with the same text. Fig.7 shows that the correlation changes with respect to the different number of references. In the case that "Optimal Similarities" metric is chosen, the correlation gets better, when the number of references increases. In theory, as long as there are enough references of one utterance which include all of its prosody styles in the database, we are always able to find its standard reference. However, when the number of references is more than six, the correlation increases slightly, which indicates that six references basically contain all of one utterance's prosody styles.

In Section 3.3 we propose two methods to measure the prosody similarities between the utterance and its references. The first metric is "Optimal Similarities" (OS). In Fig.7, based on such metric, the correlation drops rapidly when the number of reference decreases. The second metric is "Weighted-Similarities" (WS). Based on the metric, the correlation degrades placidly when the number of reference reduces. When the number of references is less than four, the correlation drops rapidly. As a whole, when the number of reference is more than five, the correlation of "Weighted-Similarities" is much lower than that of "Optimal Similarities".

5. Conclusion

Prosody plays a very important role in speech communication. In the paper, firstly detailed analysis of prosody variation of inter-speakers is given from tone, intonation and rhythm. The range of tone variation is wider than that of intonation and rhythm variation. The average tone correlation of inter-speakers is 0.7, while both intonation and rhythm are 0.81. Then based on these analyses, the method for automatic prosody evaluation is proposed. Evaluation on the database shows that the method achieves the good correlation with human scoring. However, the method proposed here is dependent on the native speech of each tested utterance, which limits the generalization for CALL systems. Future work can be focused on how to apply the conclusion to text-dependent automatic prosody evaluation.

6. Acknowledgement

The work was supported by the National Natural Science Foundation of China (No. 60575032, No. 70611120555) and 863 Programs (No. 2006AA01Z138, No. 2006AA01Z194).

7. References

[1] Cao Jianfen, 2004. Intonation Structure of Spoken Chineses: University and Specificity. *Report of Phonetic Research*, 31-38.

[2] H.Franco, L.Neumeyer, V.Digalakis, O.Ronen, 2000. Combination of Machine Scores for Automatic Grading of Pronunciation Quality. *Speech Communication*, 121-130.

[3] Jiang-Chun Chen, Jyh-Shing Roger Jang, Jun-Yi Li and Ming-Chun Wu, 2004. Automatic Pronunciaton Assessment for Mandarin Speech. *Proc. Int. Conf. on Multimedia And Expro*, Taiwan.

[4] J.Zheng, CH.Huang, M.Chu, F.K. Soong, W.P.Wei, 2007 .Generalized Segment Posterior Probability for Automatic Mandarin Pronunciation Evaluation, *Proc.ICASSP*. 201-204.

[5] L.Neumeyer, H.Franio, V.Digalakis, M.Weintraub, 2000. Automatic Scoring of Pronunciation Quality. *Speech Communication*, 83-93.

[6] Min Chu, Honghui. Dong, Jianhua. Tao, 2006. A Perceptual Study on Variablity in Break Allocation within Chinese. *3th International Conference on Speech Prosody 2006*, Dresden, Germany.

[7] Min Chu, Yong Zhao, Eric Chang, 2006. Modeling stylized invariance and local variability of prosody in text-to-speech synthesis. *Speech Communication.* 716-726.

[8] S.M. Witt, 1999. Use of Speech Recognition in Computer-assisted Language Learning. *PhD thesis, Cambridge*.

[9] W.Chu, S. S.Keerthi, 2005. New Approaches to Support Vector Ordinal Regression, *Proc.ICML*, 145-152.

[10] Y.Tian, J.L.Zhou, M.Chu, E.Chang, 2004. Tone Recognition with Fractionized Models and Outlined Features, *Proc. ICASSP*, 105-108.

[11] Huibin Jia, Jianhua Tao, 2007. Automatic Prosody Evaluation of Mandarin Speech. COCOSDA 2007. (accepted)

[12] http://www.gatsby.ucl.ac.uk/%7Echuwei/svor.htm