Combining Prosodic and Spectral Features for Mandarin Intonation Recognition

Wei Bao[1,2] Ya Li[2], Mingliang Gu[1], Jianhua Tao[2], Linlin Chao[2], Shanfeng Liu[2]

[1]Institute of Linguistic Sciences, Jiangsu Normal University, Xuzhou
[2]Institute of Automation, Chinese Academy of Sciences, Beijing
junubw@163.com,mlgu@jsnu.edu.cn,{yli,jhtao,linlin.chao,sfliu}@nlpr.ia.ac.cn

Abstract

In this paper, a feature set for Chinese Mandarin intonation is addressed. We first analyze the prosodic features, which make significant contributions to distinguishing between Chinese Mandarin interrogative intonation and declarative intonation. Then make a fusion between prosodic features and Mel Frequency Cestrum Coefficient (MFCC) feature. Our study is performed on a Mandarin question corpus, which contains large amount and various types of interrogative sentences. Experimental results indicate that the combination of prosodic features and MFCC feature achieves 77.09% classification accuracy, a 16% improvement over that of prosodic features alone. The proposed feature set performs well in the condition of speaker-independent.

Index Terms: Mandarin intonation recognition, prosodic features, feature fusion

1. Introduction

According to the Strong Universalist Hypothesis [1], in non-tonal language interrogative intonation has a rising end whereas declarative intonation has a failing end. However, a falling end in Chinese Mandarin may be a declarative intonation, which manifests in the interaction of tone and intonation. Many researchers have paid more attentions in Chinese intonation study due to its complexity in recent years.

Zhao made the initial contribution to Mandarin intonation study. He put forward 'small ripples riding on large waves' theory, which meant in Mandarin the relationship between tone and intonation was a kind of "algebraic sum". Since Zhao's research on tone and intonation, the study of intonation became a hot topic.

Lin [2][3] claimed that 'boundary tone' was necessary for intonation recognition and indicated that in yes-no questions information about question and statement was mainly carried by the last one syllable or last two syllables. Sometimes it was carried by the first syllable of the sentence. Shen [4][5] proposed that the top line of the interrogative sentences was falling gradually and the base line was rising.

Different from the perceptual experiments above, Yuan [6] built F0 curve to model the differences between interrogative and declarative intonation, showing that interrogative intonation had a higher F0 contour than declarative intonation, the last syllable had careful intonation and wider pitch swings in question sentences.

Moreover, Jiang, Cai [7][8] copied the acoustic features from question utterances to the corresponding statement ones and got the synthetic speech as the perceptual materials. The experiments results showed that pitch feature taking the most important role in Mandarin intonation recognition, and energy and duration features could compensate to a certain extent. Agreed with Cai, Wu [9] indicated duration of the last three syllables in the question intonation were shorter than that in statement intonation.

Most studies on Chinese Mandarin intonation focused on a small number of typical question and statement sentences, which had the same prosodic structure and tone combination, and relied heavily on prosodic features alone. Then draw conclusions from auditory impressions or instrumental statistics. However, the question and statement sentences in real life are variable. So a small-sample data for experiments is not universal and robust enough.

To our best knowledge, feature fusion method has not been attempted to Chinese Mandarin intonation recognition. In this paper, we proposed a good feature set which combined prosodic features and MFCC feature, to distinguish the interrogative intonation from the declarative intonation. Furthermore, we attempt at performing the experiments in a large-scale corpus. In short, our aims are twofold, a comparison of the prosody alone with feature fusion method and an evaluation of the proposed feature set in a large corpus. Results show that the proposed feature set is able to distinguish the question intonation and statement intonation well. Such method is of great interest to all research communities focusing on better performance of speech synthesis and the interaction between human and computer.

The paper is organized as follows. In the next section, a brief introduction of the database for this paper is given. In Section 3 we describe the feature set proposed in our work. Experiments and results are presented in Section 4. Finally, the conclusion is in Section 5.

2. Datasets

CASIA Mandarin question intonation corpus is collected by Institute of Automation, Chinese Academy of Sciences. It contains a total of 2356 interrogative utterances of variable content spoken by two male and two female native Mandarin speakers. 589 parallel transcriptions are selected. After the corpus was recorded, each utterance is segmented at the syllable level and then checked manually, which is good for studying Mandarin intonation based on the unit of syllable. Finally, information about syllables and tone of each syllable are reserved in tag files that provided with the wav files in the database. Recording are made with 16-bit precision and at a sampling rate of 16kHZ.

According to Shao [10], Mandarin interrogative sentences can be divided into five types. Li [11] forced on two types of the interrogative sentences: ending with "ma" or not. In this paper we select a widely accepted method. Interrogative sentences can be mainly classified into four types: Yes or No questions, Wh-questions, V bu V questions, alternative

978-1-4799-4219-0/14/$31.00

questions, as shown in Table 1. The sentences in our corpus cover these four types in a balanced way.

The declarative sentences in this work contain 1200 utterances produced by four speakers (two male speakers and two female speakers), which is a subset of CASIA emotion corpus. 300 parallel transcriptions are selected. Each utterance segmented at the syllable level too. Recording are made with 16-bit precision and at a sampling rate of 16kHZ. Two parts share no speakers and texts.

Table 1. *Four Types of Question Sentences.*

Types of Question Sentences
1. Yes or No questions
e.g. Do you believe Father Christmas?
2. Wh-questions
e.g. What kind of engine is mounted on the plane?
3. V bu V questions
e.g. You say that small pot, isn't it?
4. Alternative questions
e.g. Is it Europe's time or local time?

3. Features

In this section, we firstly elaborate the candidate prosodic feature set based on previous researches. Then make a fusion in feature level to study the roles of prosodic features and MFCC feature in Mandarin intonation recognition task.

3.1. Prosodic features

Prosodic features manifest itself in many aspects, especially in pitch, energy and duration. Question intonation gives an impression of being remarkable at the end of an utterance. In addition, syllables at the end of question intonation are in some sense perceptually longer than in the statement intonation, so we expect question and statement intonation to show significant differences in the first syllable and the last two syllables of the utterance. Based on the literature above, we think a higher energy and pitch contour is a mark of interrogative intonation.

Table 2. Original prosodic features for experiments.

Pitch	• max, min, mean of the first syllable • max, min, mean of the penultimate syllable • max, min, mean of the last syllable • top line, base line • vice-top line, vice-base line • max, min of the whole utterance • pitch range the whole utterance
Duration	• duration of the penultimate syllable • duration of the last syllable
Energy	• energy of the penultimate syllable • energy of the last syllable • energy of the whole utterance

All of the prosodic features were computed based on frames and then got common statistics considering a syllable as a unit. To eliminate differences over speakers, prosodic features were normalized with respect to the mean and standard deviation for different speaker. In total, 21 original prosodic features were extracted firstly as listed in Table 2.

3.1.1. Pitch

Guided by the previous approaches [4], we expected the rising and falling behavior of pitch contours were good cues to distinguish question intonation from statement intonation. So we calculated the relative values of pitch contour to eliminate the difference from duration of each sentence, such as top line, base line [12], which were defined as follows.

$$topline = (pitch_{first\_max} - pitch_{last\_max}) / duration \quad (1)$$

$$baseline = (pitch_{first\_min} - pitch_{last\_min}) / duration \quad (2)$$

Take the neutral tone in the end of the sentences into consideration, we computed the vice-top line and the vice-base line, which replaced the last syllable by the penultimate syllable in the formula above. To capture overall pitch range, maximum and minimum pitch value were calculated in an utterance excluding unvoiced frames.

In addition to these global pitch features above, we also computed the pitch value of the first, penultimate and the last syllable. Three common statistics (max, min, mean) were got to capture the local pitch range. Lin [2] showed that the last syllable strengthened the adjacent syllable in pitch. To address this difference, we calculate the square of the difference of the mean values of last two syllables, which is also selected as a candidate feature.

The pitch feature is processed by two principles as following:

1. Pitch values 50HZ are removed;
2. Pitch values are transformed to a log scale;

These two processing are done to better match with human perception.

3.1.2. Energy and Duration

For the energy and duration feature, we calculated three new features based on the original prosodic feature set. Due to the length of utterance was different, we calculated the relative duration instead of the absolute duration to eliminate the difference.

Table 3. *New Features Computed from Energy and Duration Features.*

Feature Name	**Description**
Energy$_{\_diff\_last2}$	Difference of energy value between the last two syllables of the utterance.
Duration$_{\_diff\_last2}$	Difference of duration value between the last two syllables of the utterance.
Duration$_{\_ratio\_last2}$	Ratio calculated from the last two syllables of the utterance.

Guided by [6], we thought that the strength of the energy of the interrogative intonation was higher than that of declarative intonation. So we expected that the interaction between the last two syllables in question intonation was a discriminating feature. In experiments we included two types of energy features, the energy value of the whole utterance and the last two syllables. To keep consistent with the pitch feature,

we only reserved the energy value of those frames, the pitch value in which is greater than 50HZ.

Last syllable of the interrogative sentence gave an impression of having a longer duration [13]. Therefore, duration of the last syllable was calculated. Wu [7] showed that question marker at the end of the utterance reduced the duration of adjacent syllable and longer of the last syllable was, more interrogative information it carries. To address this possibility, we got the duration of last two syllables firstly, then computed the ratio between them, and then calculated square of the ratio to enlarge this difference.

3.2. Spectral features

The cepstral coefficient is reported to be robust in different pattern speech recognition task. Especially, Mel-Frequency Cestrum Coefficient approximates human auditory systems more closely than the linearly-spaced frequency bands used in normal cepstrum. As a result, MFCC has been widely used for speech recognition.

As in typical speech recognition we extracted a 13 dimensional MFCC feature vector, which consists of 12 MFCC and a normalized energy. Common statistic was computed such as mean. In order to keep consistent with prosodic features and exclude the influence of the other syllables, we only extracted MFCC features from the first, the penultimate and the last syllable. 39 MFCC features were extracted from three syllables in total. To remove the redundant information, PCA (Principal Component Analysis) algorithm was used to reduce feature dimension and 95% cumulate contribution rate was used in the experiments. Finally 29 MFCC features were reserved.

4. Experiments and results

In this section, experiments will be conducted on the CASIA question corpus to evaluate the performance of our proposed feature set in speaker-independent Mandarin intonation recognition. Our proposed selected feature sets were compared to each other and support vector machines (SVM) was adopted as classifier. The kernel parameter c and g were computed via a grid search on a base-2 logarithmic scale.

To guarantee speaker-independent, the original corpus is split into two disjoint parts, three speakers for training, and the left one for testing. Two parts share no speakers. Finally performance is then taken as the average of the performance achieved in each speaker. Two experiments were conducted to evaluate the performance of the proposed feature set: (1) prosodic features only, (2) combined proposed prosodic features with MFCC feature.

4.1. Exp1. Using prosodic features only

Table 2 shows classification accuracy of each speaker using prosodic feature alone listed in Section 3. We can see that the classification accuracy of No.2 and No.3 is lower than the other two speakers and No.2 is slightly lower than No.3. We think there remains a possibility that different manner and habit has an important influence on each speaker.

So we conducted a small-scale perception experiment on whether the Mandarin interrogative intonation was strong or not. Two native Chinese speakers were asked to do perception experiments. The materials for the experiments consisted of two parts, one was the interrogative sentences misjudged in classification experiment, another part was the correct sentences selected by random. 40 speech samples were in total.

Results showed that interrogative intonation of No.1 speaker and No.4 speaker could be recognized easily. However, 75% of interrogative intonation sentences of No.2 and No.3 speakers were weaker, and some utterances were harder to be recognized. Among these interrogative utterances, 80% of the utterances were Yes or No question.

We thought the primary cause was that when speakers were recording the interrogative intonation utterances, we asked for natural revelation instead of performing exaggeratedly. In addition, Yes or No interrogative intonation sentences had the same structure with statement sentences. Due to the speakers' different habits, many utterances were harder to be distinguished from the declarative utterances.

In addition, question intonation only can be identified if listeners actually hear the question features. As 'small ripples riding on large waves' theory shows, in Mandarin the relationship between intonation tone and intonation is a kind of "algebraic sum". When the question features and the tonal features make confliction, it is hard for people to identify the question intonation. Both the results of the perceptual experiments and the previous researches are consistent with our classification experiment results.

Table 2. *Accuracy of each speaker using prosodic features only.*

Speaker	No.1	No.2	No.3	No.4
Accuracy	**64.94%**	55.69%	57.84%	**65.96%**

4.2. Exp2. Combining prosodic features and MFCC

As the flowchart in Figure 1 showed, we extracted prosodic features and MFCC feature respectively, then reduced the dimension of MFCC feature and got 95% principle components finally. Then made a fusion in feature level: combining the prosodic feature set with MFCC. Finally, SVM is used to classify the combining feature vectors.

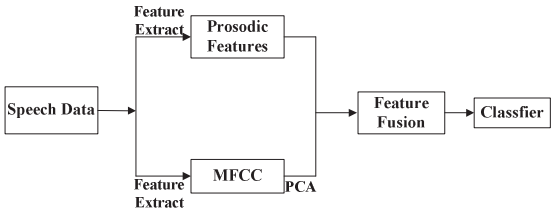

Figure 1: *Flowchart of the feature fusion method*

Figure 2 presents the comparison between using prosodic features alone and feature fusion method. In Figure 2, "P" in the labels denotes the prosodic features only method and "F1" denotes the fusion with original MFCC features method. "F2" is the result when make a fusion between prosodic features with the reduced-dimension MFCC features. Labels on the horizontal axis represent the classification accuracy testing on the specified speaker and "mean" denotes the mean accuracy of the four speakers.

Clearly, a better classification rate is obtained using the fusion with reduced-dimension MFCC feature method. An

average 16% improvement is got than using prosodic features only.

According to Figure 2, conclusions can be made as follows:

- "P" method is the lowest among the three methods;
- "F1" method gets about 2% improvement than "P" method;
- "F2" method achieves the highest classification accuracy for No.1, No.2 and No.3 speaker, but fail to make improvement for No.4;

This can be attributed to two main causes. First, combination with MFCC feature improves the classification accuracy of prosodic features only to some extent. Secondly, principal component analysis algorithm is used to reduce the dimension of the MFCC feature and 95% cumulate contribution rate is selected finally, which can remove the redundant information. For No.4 speaker, "F2" method do not improve the classification accuracy, this can be accounted for that 95% principle components of MFCC feature lose some important information, which is necessary.

In addition, by analyzing the result of No.2 speaker, we find that different from the other three speakers, interrogative sentences from No.2 speakers were easier to be recognized as declarative sentences. This is mainly caused by the data from No. 2 speaker is not standardized enough and the weight of SVM classifier is set according to the whole sample rather than each speaker.

Clearly, while prosodic feature set do not perform so well on its own, when combined with the MFCC feature, the combining feature set outperform prosodic features alone.

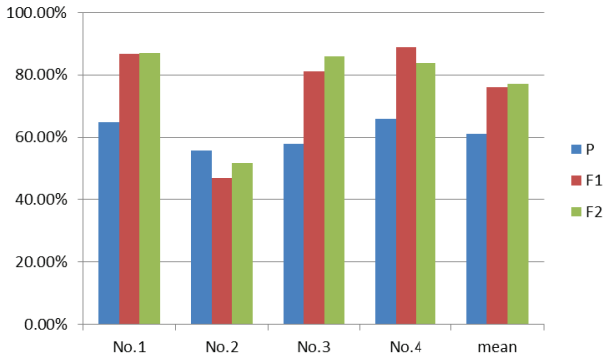

Figure 2: *Performance comparison among prosodic features only, feature fusion method and fusion with reduced-dimension MFCC.*

5. Conclusions

This paper focuses on exploring another, relatively untapped method for automatic Chinese Mandarin intonation recognition. Firstly we analyzed the typical prosodic features for distinguishing Mandarin interrogative intonation and declarative intonation, and then got a prosodic feature set building on the previous researches. Speaker-independent classification experiments were conducted to evaluate the performance of the proposed prosodic feature set. A perceptual experiment was also conducted to analyze the reason why the accuracy of No. 2 and No. 3 speaker was lower than the other two speakers.

A 16% improvement was achieved after combining with spectral feature. To evaluate the proposed feature set in a robust condition, we guaranteed that experiments were conducted in the condition of speaker independent and the corpus was large. Experimental results suggested that our feature set was efficient enough to handle larger amounts of interrogative and declarative speech materials.

Future work includes decreasing the dimension of the proposed feature set to satisfy practical demand. In addition, textual feature is considered to be made a fusion with the proposed feature set. Some other classifiers are considered to make contrast experiments with SVM.

6. Acknowledgements

We thank Institute of Automation, Chinese Academy of Sciences for providing corpus, and help. This work is mainly supported by Program Granted for Scientific Innovation Research of College Graduate in Jiangsu Province (No. CXZZ13_0965), and the National Natural Science Foundation of China (NSFC) (No.61273288, No.61233009, No.61203258, No.61305003, No. 61332017, and No.61375027), and partly supported by the Major Program for the National Social Science Fund of China (13&ZD189).

7. References

[1] D. R. Ladd, "On Intonational Universals", In Myers, T. et al. (Ed.) The cognitive Representation of Speech, Amsterdam: North Holland Publishing, 1981.

[2] M. C. Lin, "Chinese Boundary Tones and Their Independent Role: with an Additional Investigation of the Universals and Peculiarities of Chinese and English Intonation", Proceedings of the Sixth National Conference on Modern Phonetics, 2003.

[3] M. C. Lin, "Interrogative vs. declarative and the boundary tone in Standard Chinese", Studies of the Chinese Language, Vol. 4, pp. 364-376, 2006.

[4] J. Shen, "Intonation Structure and Intonation Types of Chinese", Fangyan, Vol.4, pp. 221-228, 1994.

[5] J. Shen, "A Preliminary Study of Chinese Intonational Model", Yu Wen Yan Jiu, Vol.4, pp. 16-24, 1992.

[6] J. H. Yuan, C. Shih, G. P. Kochanski, "Comparison of Declarative and Interrogative Intonation in Chinese", Speech Prosody, pp.711-714, 2002.

[7] D. N. Jiang, L. H. Cai, "The Study on the Acoustic Features of Chinese Interrogative Mood", Proceedings of the Sixth National Conference on Modern Phonetics, pp. 186-191, 2003.

[8] Y. X. Wang, J. Jia, L. H. Cai, "Analysis of Chinese Interrogative Intonation and its Synthesis in HMM-Based Synthesis System", Proceedings of the Internet Computing and Information Services, pp. 343-346, 2011.

[9] Y. H. Wu, J. H. Tao, J. L. Lu, "Prosodic Analysis of Chinese Mandarin Intonation", Proceedings of the Seventh National Conference on Chinese Phonetics, 2006.

[10] J. M. Shao, "The Study of Modern Chinese Interrogative Sentences", Shanghai: East-China Normal University Press, 1996.

[11] Z. X. Li, P. Martin, G. Boulakia, "Tones and Intonation in Declarative and Interrogative Sentences in Mandarin", Proceedings of the International Symposium on Tonal Aspects of Language, 2004.

[12] M. L. Wang, M. C. Lin, "Declination in Mandarin Natural Discourse", Proceedings of the Sixth National Conference on Modern Phonetics, 2003.

[13] J. H. Yuan, D. Jurafsky, "Detection of Question in Chinese Conversational Speech", Automatic Speech Recognition and Understanding, pp. 47-52, 2005.