2010 3rd International Conference on Computer and Electrical Engineering (ICCEE 2010)
IPCSIT vol. 53 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V53.No.2.28

Chinese Intonation Classification Using Support Vector Machines

Jingyang XU[+], Dengfeng KE, Lei JIA

Digital Media Content Technology Research Center, Institute of Automation, Chinese Academy of Sciences

Beijing, China

Abstract—In conventional speech recognition system, only a plain text is presented as the final result, and all acoustic information of speech are cutoff. The aim of this publication is to add intonation information to traditional output of speech recognition engine, which is believed to reflect the emotion and intention of speaker. In this paper, we propose a robust approach to classify several kinds of intonations, e.g. declara-tive, interrogative, exclamatory, etc. Since it is still an open question on how to describe intonations, different kinds of features are investigated here to choose the most effective features for intonations classification. Support Vector Machine (SVM) is used as the classifier to perform the task of feature selection and combination. In our experiment, we address the speech recognition based methods, and use recognized results replace the transcribed text. Our goal is to simulate intonation classification in the real speech recognition. The speech materials used in this experiment were well designed includes three intonations, total about 4700 sentences. Experimental results show that our system can achieves the accuracy of (84.13%) for the task of three types of Chinese intonation classification.

Keywords-Chinese intonation; SVM; speech recognition; intonation classification

1. Introduction

Speech can convey not only literal information, but also the mood and emotion of a speaker. Generally, intonation refers to the variations in the pitch of a speaker's voice used to convey or alter meaning[6]. And it is contained in speech, but could not been recognized by traditional automatic speech recognition (ASR) system. The automatic classifica-tion of intonation can present discourse structures into the dialog system that aim to achieve natural language under-standing. Even for the sentences with completely same text content, different intonation may cause different under-standing during conversion.

Several studies, [1]-[6],[14],[15], have investigated different approaches for intonation classification. Many researches indicate that there is the most important distinctness in the end of sentence between declarative and interrogative modal. [1],[2] have shown a tone-dependent mechanism of question intonation at the sentence-final position. [3],[4]'s results also indicates that difference between statement and question intonations in Mandarin is manifested by an increasing departure from a common starting point toward the end of the sentence. Unlike declarative and interrogative, rare research on intonation can drive exclamatory into experiment. Some researchers point that the main feature of exclamatory is strong stress and wide tonal range of utterance. [15] utilizes Decision Trees to classify Statement and Question intonation. These methods are limited in the small scale dataset and vulnerable to noise data. They almost are based on comparative methods, and did not provide a practical framework of automatic intonation classification for speech recognition.

In this paper, we exploit a novel approach to three type intonations classification with large scale dimension feature vector hybrid speech recognition, in which we use the recognized text replace the certain

[+] Corresponding author.
E-mail address: jingyangxu@hitic.ia.ac.cn

text. SVM classifier can automatically control generalization and parameterization as part of the overall optimization process, and can better solve small sample learning problems and sparse data. Therefore, SVM classifier is employed in our classification system. For feature effectiveness, different features are compared, combined and selected to build a mixed prosody features for robust intonation classification. In this paper, we investigated prosody features composed by F0, energy, duration, and F0 curve, etc. Finally, we compare the classification accuracy of three types of Chinese intonation each other. The experimental results show that our method could achieves a good accuracy in Chinese intonation recognition.

The outline of this paper is as follows. In Section II, describes an overview of our classification system; Section III reminds the features used in Chinese intonation classification; Section IV details the data set and presents experimental results of our system; Section V discusses the implication of the results.

2. Methods

2.1. Intonation classification with SVM

Since support vector machine (SVM) [7] is one example of a classifier that estimates decision surfaces directly rather than modeling a probability distribution across the training data. SVM have demonstrated good performance on pattern recognition problems. Fig. 1 shows a typical 2-class problem in which the examples are perfectly separable using a decision region. H1 and H2 define two hyperplanes. The distance separating these hyperplanes is called the *margin*. The closet in-class and out-of-class examples lying on these two hyperplanes are called the *support vectors* [9].

An SVM [10] is a two-class classifier constructed from sums of a kernel function $K(\cdot,\cdot)$,

$$f(x) = \sum_{i=1}^{L} \alpha_i t_i K(x, x_i) + d \,, \tag{1}$$

Where the t_i are the ideal outputs, $\sum_{i=1}^{L} \alpha_i t_i = 0$, and $\alpha_i > 0$. The vectors x_i are support vectors and obtained from the training set by an optimization process [11]. The ideal outputs are either 1 or -1, depending upon whether the corresponding support vector is in class 0 or class 1, respectively. For classification, a class decision is based upon whether the value, $f(x)$, is above or below a threshold.

The kernel $K(\cdot,\cdot)$ is constrained to have certain properties (the Mercer condition), so that $K(\cdot,\cdot)$ can be expressed as

$$K(x, y) = b(x)^t b(y) \,, \tag{2}$$

where $b(x)$ is s mapping from the input space (where x lives) to a possibly infinite-dimensional SVM expansion space.

For a separable data set, SVM optimization chooses a hyperplane in the expansion space with maximum margin [10]. The data points from the training set lying on the boun-daries are the support vectors in equation (1). The focus of the SVM training process is to model the boundary between class [11].

For intonation classification, the goal is to determine the intonation of an utterance from a set of known intonations (De, In, Ex). Generally, the SVM is a 2-class classifier. In our system, intonation is a multiclass classification, so we use a *one vs. one* strategy for intonation classification. Therefore, the well-known SVM toolkit *LIBSVM2.90* [8] is employed in the system. The kernel function used in our system is the Radial Basis Function (RBF) kernel,

$$K(x, y) = \exp(-\frac{\|x - y\|^2}{\sigma^2}) \,, \tag{3}$$

where σ is found with cross-validation.

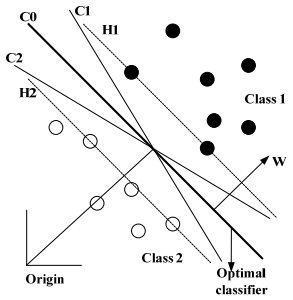

Fig.1. 2-class hyperplane classifier example

2.2. Feature Extraction System

In this paper, we only adopt three intonations recognition in our experiment. In intonation classification system, the task is to identify the utterance's intonation type. Fig. 2 is the feature extraction flowchart of intonation features. The feature set used in these experiments is derived from the F0 and energy contours of dialog act [12]. The F0 extraction algorithm uses the ESPS “get_f0” algorithm [13]. In training and testing terms, we first use ASR to align the phone or syllable boundaries of each utterance. Meanwhile, F0 and energy feature based on frames is extracted. After force alignment to the speech, we get the boundaries of syllable. According to syllable segmentation, the original pitch curve is split into two things: the F0 curve and the duration curve. Using syllable segmentation, F0 curve of entire utterance also extracts from the raw frames based F0. Then we resample this curve into 10 points curve. When these all be done, mixed the feature and input it to SVM. After gaining the intonation, we can token the recognized sentence with intonation for speech recognition results.

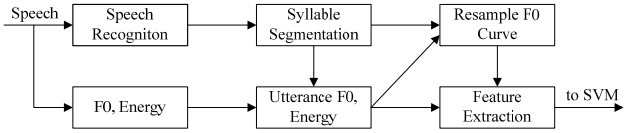

Fig.2. Feature extraction

3. Intonation Feartures

In this paper, we used the acoustic features mentioned by Shriberg et. al.[12] that made use of pitch, duration, intensity, etc. In this paper, we also introduce F0 curve as our input features. A brief description of our features can be found in Table 1. Each feature is scaled to between -1 and 1 based on the minimum and maximum values of the feature among training examples.

3.1. Pitch Feature

F0 features included both raw values (obtained from ESPS “get_f0”) and values from a linear regression (least-squares fit) to the frame-level F0 values.

Most researchers consider the F0 as the most importance feature in intonation recognition. We also first test the F0 feature set. Some Chinese linguists hold the argument that the declarative and interrogative intonation is impressed on the boundary syllable and the mainly feature of exclamatory is strong stress and wide tonal range of utterance. So we use the maximum F0 of the whole utterance and the range of F0 of utterance as the discriminative features, respectively label as Max and Range in Table 1. Moreover, F0 values were normalized on a log scale.

3.2. Duration Feature

Duration was expected to be a good cue for discriminating intonation. Generally, the three intonations have the discriminative duration in the end of sentence. For example, statement has the more evenly duration of each syllable; question has the longest duration in the end of utterance; exclamatory has the longest duration at the stress word. So we test the duration of final syllable, and labeled as LastDur in Table 1.

3.3. Intensity Feature

Unlike the most researchers use the utterance energy, because sub-space energy is one of the most important stress features, we compute sub-space energy feature to capture the gap of energy in utterance. Therefore, entire energy is sepa-rated into six sub-spaces. In each sub-space, we calculate the divergence between maximum of sub-space energy and min-imum. Thus we obtain a six dimension vector of energy (marked as En in Table 1). Similarly, by dividing sentences into two parts: former sentence and latter sentence, we calculate sub-space energy difference of former and latter (En1 and En2 in Table1).

3.4. F0 Curve Feature

F0 curve is considered as the tendency of pitch over tonal. In our experiment, we first get the maximum, minimum, and mean of every syllable's pitch. Then maximum F0 curve is resample from the maximum pitch into 10 points. And F0 slop in an utterance is another important feature for intonation. We extraction three F0 slop from whole utterance, the first dimension is the former utterance's F0 slop, the second is the latter's, and the last dimension is the overall utterance's slop.

TABLE I. BRIEF DESCIPTION OF FEATURES

Feature Sets	***Feature and Brief Description***
F0	LastPitch: mean of pitch of last syllable Range: tonal range, F0 Max subtract Min Ave: mean of pitch of whole sentence Max: maximum of F0 Min: minimum of F0 Range12: 2 dimension vector of former and latter tonal range Ave12: 2 dimension vector of former and latter mean of pitch RangeRate: ratio of average of tonal range of former and latter RangeDiff: latter tonal range subtract the former Max12: 2 dimension vector of former and latter max F0 Min12: 2 dimension vector of former and latter min F0 MaxDiff: max F0 of latter subtract the former AveDiff: min F0 of latter subtract the former AveRate: mean F0 of latter subtract the former
Duration	LastDur: duration of final syllable Dur12: 2 dimension vector of normalization of duration of former and latter sentence DurRate: ratio of mean of duration of former and latter sentence
Intensity	En: 6 dimension vector of sub-space energy gap En1: 6 dimension vector of sub-space energy gap of former sentence En2: 6 dimension vector of sub-space energy gap of latter sentence
F0 Curve	Max_Curve: maximum F0 resample Min_Curve: minimum F0 resample Mean_Curve: mean F0 resample Slop: 3 dimension of F0 slop abstract

4. Experiment

4.1. Data Set

The speech materials used in this experiment were well designed includes three intonations: declarative (De), interro-gative (In), exclamatory (Ex). The intonation speech data-base is consisted by 37 males and 40 females. Each person records about 62 sentences covering three types of intonations. The reading text includes about 124 independent sentences. We separate the speech data to training set and test set, according to the text. To obtain the data which just stands for the suprasegmental feature, we design the training and test set text with these rules: first, the tone of last syllable in the training set and the test set is different; second, if is it is the same, we ensure the last Chinese character is not the same word. Table 2 shows the number of sentence in training and test set, and types of intonations.

TABLE II. DETAIL OF TRAINING SET AND TEST SET

	Declarative (De)	***Interrogative (In)***	***Exclamatory (Ex)***

	Declarative (De)	*Interrogative (In)*	*Exclamatory (Ex)*
Train	1152	1142	1062
Test	353	584	384

4.2. Results

Since our first goal is to recognize intonation type, we also interested in feature importance. So we take the strategy that we gradually add feature set to our classification system, in order to compare different feature how it make the classification correct rate (CCR) performance different. Under this strategy, we first test only using each feature set. Table 3-6, shown the CCR of F0 set, Duration set, Intensity Set and F0 Curve set.

TABLE III. CCR OF F0 FEATURE SET

Feature	*De/In/Ex*	*De/In*	*De/Ex*	*In/Ex*
LastPitch	42.41	60.98	53.10	57.20
Range	39.68	60.55	53.10	53.58
Ave	38.40	62.79	51.88	52.34
Max	**49.98**	71.01	56.36	**67.63**
Min	36.12	60.12	53.10	48.93
Range12	45.97	**73.89**	**63.96**	54.82
Ave12	33.78	64.71	44.96	43.67
RangeRate	36.73	61.83	53.78	49.76
RangeDiff	41.58	57.24	57.04	55.65
Max12	45.51	61.54	53.10	64.68
Min12	37.87	61.19	45.51	48.00
MaxDiff	45.21	63.86	53.10	61.43
AveDiff	34.31	63.33	48.22	46.45
AveRate	33.48	63.33	48.76	45.32

TABLE IV. CCR OF DURATION SET

Feature	*De/In/Ex*	*De/In*	*De/Ex*	*In/Ex*
LastDur	57.77	**88.4**	49.98	**78.47**
Dur12	**59.28**	75.81	84.17	69.69
DurRate	55.57	72.93	**86.88**	66.18

TABLE V. CCR OF INTENSITY SET

Feature	*De/In/Ex*	*De/In*	*De/Ex*	*In/Ex*
En	**54.22**	**74.21**	**77.26**	**64.02**
En1	36.73	52.01	61.79	60.40
En2	48.84	64.93	76.71	57.09

TABLE VI. CCR OF F0 CURVE SET

Feature	*De/In/Ex*	*De/In*	*De/Ex*	*In/Ex*
Max_Curve	52.40	84.99	**63.69**	**65.05**
Min_Curve	**58.31**	**89.47**	61.11	60.09
Mean_Curve	55.05	87.55	48.22	61.64
Slop	43.62	63.33	57.99	59.37

From Table (3-6), we can obviously get that F0 feature set is gets the worst performance, and duration feature set has the best performance in classification. This result keeps with other researchers' investigation that duration was expected to be a good cue for discriminating Statements and Questions [12]. For former sentence and latter sentence performance, the results confirm that there is the most important distinctness in the end of sentence.

After testing four different feature sets, we then select the good performance features to build a set of new features, so-called optimal features. In F0 set, according to the performance and acoustic knowledge, we choose Range, Ave, Max, Min, Range12, and Ave12 as the F0 selected set (F0 S). And choose En, and En2 as the Intensity selected set (Intensity S). Using whole Duration and F0 Curve set in feature set. Table 7 lists the results of different mixed feature set.

TABLE VII. CCR OF MIX FEATURE SET

Feature Set	*De/In/Ex*	*De/In*	*De/Ex*	*In/Ex*
F0 +Duration	65.80	77.41	79.29	82.09
F0 S + Duration	70.19	80.94	81.19	84.99
F0 S + Duration + Intensity S	73.82	83.07	92.59	85.81
F0 S + Duration + Intensity S + F0 Curve	**84.13**	**86.10**	**93.35**	**92.46**

Table 7 shows the CCR of different mixed features. Mixing F0 selected set achieve a better performance than mixing whole F0 set, this may be caused by applying duplicated feature.

5. Conclusion

This study use pitch, duration, energy and F0 curve feature to classification intonation. We first take Exclama-tory modal into account. Experiment results reveal that duration is the most important feature for intonation, and pitch is the worst feature for it. This may be caused by the F0 algorithm. After selecting and mixing these acoustic features, we achieve a satisfactory classification correct rate (84.13%) in three types of Chinese intonation classification. This classification has a good performance in speech recognition results. Although we have got significant classification accuracy, there is still an open issue on intonation or modal recognition. Future work should be investigated the intonation recognition on the large speech data and spontaneous conversion.

6. References

[1] J. K-Y. Ma, V. Ciocca, and T. L. Whitehill, "Acoustic Cues for the Perception of Intonation in Cantonese," Proc. INTERSPEECH 2008, Brisbane, Australia, Sept. 2008, pp. 520-523.

[2] J. K-Y. Ma, V. Ciocca, and T. L. Whitehill, "Quantita-
tive Analysis of Intonation Patterns Produced by Cantonese Speakers with Parkinson's Disease: A Preliminary Stud y," Proc. INTERSPEE-CH 2008, Brisbane, Australia,, pp. 1749-1752.

[3] J. Yuan, "Perception of Mandarin Intonation," Proc. ISCSLP 2004, Hong Kong, P. R. China, Dec. 2004, pp. 45-48.

[4] J. Yuan, "Mechanisms of Question Intonation in Mandarin," Proc. ISCSLP 2006, Singapore, Dec. 2006, pp. 19-30.

[5] J. Yuan, and C. Shih, "Confusability of Chinese Intonation," Proc. Speech Prosody 2004, Nara, Japan, March 2004.

[6] D. Ke, B. Xu, "Chinese intonation assessment using SEV features," Proc. ICASSP 2009, pp. 4853-4856.

[7] V.N. Vapnik, Statistical Learning Theory, John Wiley, New York, New York, USA, 1998.

[8] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[9] A. Ganapathiraju, J.E.Hamaker, J.Picone, “Applications of Support Vector Machines to Speech Recognition,” IEEE Trans. on Signal Processing, Conversay, USA, vol. 52, NO. 8, 2004, pp. 2348-2355, August 2004.

[10] N. Cristianini, J. Shawe-Taylor, “Support Vector Machines,” Cambridge University Press, Cambridge, 2000.

[11] W. M. Campbell, D. E. Sturim, D.A. Reynolds, and A. Solomonoff, “SVM Based Speaker Verification Using A GMM Supervector Kernel and NAP Vatiability Compensation,” Proc. ICASSP 2006.

[12] E. Shriberg, R. Bates, and A. Stolcke, “Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech”, In M. Swerts and J. Hirschberg (eds.) Special Double Issue on Prosody and Conversation. Language and Speech 41(3-4), 1998, pp. 439-487.

[13] D. Talkin, “Robust algorithm for pitch tracking,” Speech Coding and Synthesis, pp. 497–518, 1995.

[14] Z. Su, and Z. Wang, "Affective Intonation-Modeling for Mandarin Based on PCA," computational linguistics and chinese language processing, vol. 12, No. 1, March 2007, pp. 33-48

[15] F Liu, D Surendran, Y Xu, "Classification of Statement and Question Intonations in Mandarin," Proc. Speech Prosody, Dresden, 2006.