Automatic Accent Classification Using Ensemble Methods

Fukun Bi, *Jian Yang*, *Dan Xu*

School of Information Science and Engineering,
University of Yunnan, Kunming, 650091, P. R. China
bifukun@163.com, nxryang@126.com, danxu@vip.sina.com

10.21437/Interspeech.2008-234

Abstract

Accent classification technologies directly influence the performance of the state-of-the-art speech recognition system. In this paper, we propose a novel scheme for accent classification, which uses decision-templates (DT) ensemble algorithm to combine base classifiers built on acoustic feature subsets. Different feature subsets can provide sufficient diversity among base classifiers, which is known as a necessary condition for improvement in ensemble performance. Compared with those methods of Majority Voting ensemble and Support Vector Machine, our ensemble scheme can achieve the highest performance. On the other hand, we investigate the possible reasons why ensemble systems can provide potential performance, in terms of diversity analysis. In our experiments, a native Mandarin speech corpus and a non-native multi-accent Mandarin speech corpus which contains three typical minorities' accents in Yunnan, China, are adopted.

Index Terms: accent classification, Yunnan minorities' accents, ensemble methods, diversity.

1. Introduction

Improving the performance of the state-of-the-art Mandarin speech recognition system for non-native speech remains a challenging task since the wide varieties of non-native accents. Most of the minorities in China, such as Naxi, Dai, Lisu etc., have their languages, they speak Mandarin with their native language accents. If the accents can be classified accurately, then more adapted models which considering speaker accent could be employed to increase the accuracy of Mandarin speech recognition.

There are two active research fields related to accent issues: accent adaptation through pronunciation modeling and accent classification. In this paper, we confer special attention on the second issue. Many successful methods to accent classification have been proposed, classical approaches are HMM [1][15], and Gaussian mixture model (GMM) which can avoid building model for phone or phoneme-class and achieve high-accuracy recognition [2][16]. A more recent development is the uses of Support Vector Machine (SVM) which also is demonstrate an efficient method [3]. On the other hand, some studies focus on acoustic features of accent, a comparison of two unsupervised approaches with different acoustic levels is introduced in [4]. In addition, some investigates pay their attention on the characteristics of the accent itself, for example, a content word is frequently followed by a particle to form an utterance unit with one accent component in Japanese continuous speech [5], and there are various phonetic correlates of sentence accents from those of word stress in Finnish [6].

Classifier ensemble is a popular approach to improve the performance of recognition systems. We consider the application of this idea to accent classification. It is intuitively accepted that classifiers to be combined should be *diverse* [8]. Unlike most of previous studies constructed a super-vector for accent classification by fusion feature subsets which from acoustic features of speech [1][3]. The key part of our approach is that we use decision-templates ensemble algorithm (DT) to combine base classifiers built on different acoustic feature-subsets which can provide sufficient diversity among base classifiers. In addition, it is well-known that diversity analysis is closely related to ensemble systems [7][8]. A better investigation of diversity can be expected to achieve higher performance in our ensemble scheme. However, to our knowledge, there is no relative work investigates the ensemble systems for accent classification in terms of diversity analysis. In this paper, we use diversity analysis to discuss the possible reasons why DT ensemble can provide potential performance than Majority Voting which is a common method for ensemble strategy.

The rest of this paper is organized as follows: In Section 2 we introduce an accent corpus adopted in our experiments. Section 3 discusses the extraction of acoustic features. Section 4 applies the DT algorithm in our ensemble scheme. The diversity analysis is discussed in Section 5. Experimental results are showed in Section 6. Section 7 concludes the paper.

2. Accent corpus

In our experiments, two speech corpora shown in Table 1, one native speech corpus and one non-native speech corpus, are used. The native Mandarin speech data are extracted from the Mandarin Dictation Corpora supported by *China National Hi-Tech Project 863*. The non-native Mandarin speech data are extracted from the *Linguistic Minorities Accent Mandarin Speech Corpus (LMAMSC)*, which collected by our laboratory and ever used for continuous Mandarin speech recognition for non-native speaker in our precious work [9]. These utterances of two corpora were both recorded under studio conditions with high-quality recording equipments and saved in mono wave files with 16,000Hz sample rate and 16 quantitative bits. The maximum utterance duration is 14.12s, the minimum is 1.27s, the mean duration is 3.45s, and the standard deviation is 2.18s. The speakers of non-native speech corpus are from minority areas in Yunnan and their native languages are not Chinese. The non-native accents are obvious when they speak Mandarin. For the studies here, the focus was on Mandarin

Table 1. *Speech corpus overview.*

Corpus	Accent	Speakers	Utterances
Project 863	Native	16	240
LMAMSC	Naxi	16	240
	Lisu	16	240
	Dai	16	240
Total	4	64	960

across the following three accents from LMAMSC: *Dai*, *Naxi*, and *Lisu*. Each accent corpus contains utterances of both male and female speakers (8 males and 8 females), thus final accent corpus of 960 utterances was created with 15 utterances per speaker and these utterances have different contents with each other.

3. Feature extraction

Acoustic features are widely observed to carry the most significant characteristics of accent in speech [1][2][3]. In this study, we estimated the following acoustic features: fundamental frequency (F0), short-time energy (En), first formant (F1), and Melfrequency Cepstral Coefficients (MFCCs), and calculated their correlative statistical features. However, we carried out the investigations at utterance level.

F0 was extracted using the autocorrelation algorithm. To extract F1, we found the poles of the autoregressive transfer function with the linear predictive coding (LPC) coefficient. Thus, we estimated F0, F1 and En of each frame in speech signals and connected their corresponding dots to form a raw contour respectively. Then the raw contours were smoothed with the median filter and linear filter. For each contour, we calculated the statistical features: maximum, minimum, mean, range (max-min), standard deviation, skewness, kurtosis, mean of jitter and range of jitter. The first 12 MFCCs were extracted from every frame, and the mean were calculated.

Finally, the acoustic-feature vector for each speech utterance consists of 39 features, to reduce the variability among different feature groups, we normalize these features. Table 2 shows the distribution of these statistical features.

Table 2. *Distribution of the statistical features for each utterance.*

Feature group	Number
Fundamental frequency	9
Format	9
Short-time energy	9
MFCCs	12
Total	39

4. Classification methods

It is believed that ensemble techniques have better potential for improvement on accuracy than single-based classifiers in pattern recognition systems. They are widely used in many pattern recognition tasks, also are they used successfully in speech field in recent works as in [14]. In this paper, we apply decision templates (DT) method in ensemble strategy, which was proposed by L. I. Kuncheva in [10] and was reported a simple and effective method. Especially, it could yield better performance than other similar ensemble schemes. In order to compare the performance of this ensemble scheme, single-based classifier and Majority Voting ensemble are used for comparative experiments.

4.1. Decision Templates ensemble

4.1.1. Base classifiers and its training strategies

The reasonable choice of base classifiers is fundamental to the overall performance of an ensemble scheme. We choose Support Vector Machine (SVM) with radial basis function (RBF) kernel as the base classifier. It usually shows the highest performance in previous works of speech issues as in [11] [12]. Let $\{ D_1, D_2, \ldots, D_i, \ldots, D_L \}$ be a set of base classifiers, and the number L of them is equal to that of feature subset. They were trained by the corresponding feature subsets, for example, the i-th feature subset training the base classifier D_i. Note that, one feature subset consists of a certain feature-group (as in Table 2.).

4.1.2. Estimation of decision templates

Each base classifier (trained) gets as its input a certain feature vector $\mathbf{x}_i \in \mathbf{X}$ i=1,2…L, which comes from the corresponding feature subset. Let $\{w_1, w_2, w_3, w_4\}$ be the set of class labels which represent 4 accents, we assume that the feature subsets are denoted by the same label with the utterances which they come from. The classifier outputs can be organized in a decision profile (*DP*) as following matrix.

$$DP(\mathbf{x}) = \begin{bmatrix} d_{1,1}(\mathbf{x}_1) & \ldots & d_{1,j}(\mathbf{x}_1) & \ldots & d_{1,4}(\mathbf{x}_1) \\ \ldots & & & & \\ d_{i,1}(\mathbf{x}_i) & \ldots & d_{i,j}(\mathbf{x}_i) & \ldots & d_{i,4}(\mathbf{x}_i) \\ \ldots & & & & \\ d_{L,1}(\mathbf{x}_L) & \ldots & d_{L,j}(\mathbf{x}_L) & \ldots & d_{L,4}(\mathbf{x}_L) \end{bmatrix} \quad (1)$$

Where, $d_{i,j}(\mathbf{x}_i)$ is the degree of "support" given by the base classifiers D_i, for the hypothesis the given input $\mathbf{x}_i$ comes from the i-th feature-subset which labeled w_j, i=1,2,…,L, j=1,2,3,4. "Crisp" class labels are adopted in this work based on the outputs of SVM base classifiers. For example, $d_{i,j}(\mathbf{x}_i)$=1, if D_i recognizes correctly, otherwise, $d_{i,j}(\mathbf{x}_i)$=0.

Let Z = $\{z_1, z_2, \ldots, z_m, \ldots, z_n, \ldots, z_p\}$ be the labeled data set. The i-th row of $DP(z_m)$ is evaluated by the i-th feature-subset of the utterance z_m. The decision template j (DT_j) for accent class j is the expectation of the *DP*s which are evaluated by training utterances labeled as class j.

$$DT_j = \frac{1}{N_j} \sum_{z_m \in w_j \in Z} DP(z_m), \quad j=1,2,3,4, \quad (2)$$

where N_j is the number of these training utterances. Thus we obtain 4 *DT*s denote 4 accents with all training utterances of the corpus.

4.1.3. Classification with decision templates

In the test phase, for test utterance z_n, we use its feature subsets to calculate the $DP(z_n)$. μ_j (z_n) is defined as the similar degree between the current $DP(z_n)$ and DT_j, which is calculated by the Euclidean distance.

$$\mu_j(z_n) = \sum_{i_1=1}^{4} \sum_{i_2=1}^{L} (d_{i_2,i_1}(z_n) - dt_j(i_2,i_1))^2 \quad (3)$$

Where, $dt_j(i_2,i_1)$ is the i_2, i_1-th entry in DT_j. Thus, a test utterance z_n is assigned class label w_k, when μ_k (z_n) is the smallest value among $\{\mu_1(z_n), \mu_2(z_n), \mu_3(z_n), \mu_4(z_n)\}$.

4.2. Comparative classification methods

In order to compare the performance of our ensemble scheme, single-based classifier and Majority Voting ensemble are used for comparative experiments.

- We choose SVM with RBF kernel as single-based classifier, as well as the base classifiers in our ensemble schemes. The accuracy of this classification method is calculated with different feature subsets for comparison.

- Majority Voting is one of the most popular techniques used in classifier fusion. To compare the performance of DT ensemble, we also use same manner that each base classifier gets as its input a certain acoustic feature subset for Majority Voting scheme.

5. Diversity analysis

Diversity among base classifiers is recognized as one of the important characters in classifier ensemble scheme [7][8]. Both theoretical and empirical researches have demonstrated that a good ensemble is one where the base classifiers in the ensemble are both accurate and tend to err in different parts of the instance space. One effective approach for generating an ensemble of accurate and diverse base classifiers is the use of different feature subsets [13]. In our scheme different acoustic feature subsets used to generate the base classifiers, it is possible to promote diversity and produce base classifiers that tend to err in different sub-areas of the instance space. We are interested in whether there is any connection between accuracy and diversity by using different combinations of feature subsets. For this hypothesis, we use diversity analysis to compare the diversity among base classifiers which are built on different combinations of feature subsets.

There are lots of ways to quantify the *diversity* of ensemble classifiers. In our case, we use the entropy measure (Ent), which was proposed in [8] and defined as:

$$Ent = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{L/2}\min\{l(z_j), L-l(z_j)\} \cdot \qquad (4)$$

Where, $l(z_j)$ is the number of base classifiers that can correctly recognize z_j. Correct and incorrect are two possible outputs of classifiers in this case, which are denoted respectively as 0 and 1. *Ent* is within the interval [0, 1]. The higher the *Ent* value is, the greater the diversity for ensemble classifiers.

6. Experimental results and discussion

Considering the speaker variability affects the performance of our experiments, we employ the 8-fold cross-validation technique. The whole database is equally divided into 8 sub-databases (one sub-database correspond a certain speaker). In each round, we use 7 sub-databases for training and the remainder one for test, and this process is repeated for eight times, note that in our experiment half of training sets for training base classifiers and another for estimating decision templates. Gender-independent experiments are carried out for our experiments, male and female data are considered separately. As a result, all final results, by the 8-fold cross-validation, are mean of male and female results.

6.1. Performance of single SVM

In order to investigate how effectively these acoustic cues could be used to discriminate accents, we used different feature subsets and their combinations for single SVM to explore performance (a super-vector is constructed for classification by fusion different feature subsets).

As can be seen in Figure 1, for the single feature subset, the accuracy of F0 (pitch and its correlative statistical features) is better than that of others. It is confirmed that pitch is most important cues to describing accents. It also shows energy and F1 are essential in disguising accents, while MFCC are less important. We observe that combinations of feature-subsets are outperform single feature subset on average but the improvement is limited. This is likely due to the fact that there is much redundant accent information among these acoustic features. Note that the combination with more feature subsets does not always outperform the less one. Such as, the highest performance of classification is achieved by combination with F0, En and F1, but not by combination with all acoustic features.

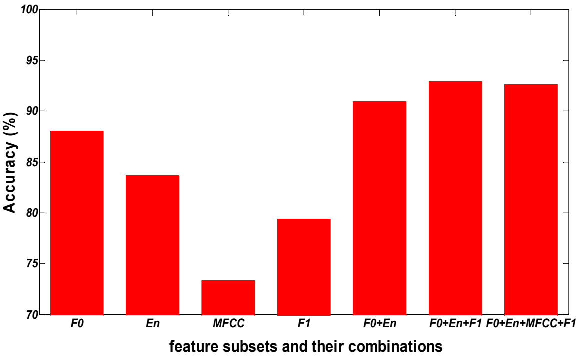

Figure 1: *Comparison of different feature subsets and their combinations for single SVM.*

6.2. Performance of ensemble classifiers

To compare the performance of our ensemble scheme, single SVM and Majority Voting ensemble are used for comparative experiments. Figure. 2 shows the accuracy of two ensemble schemes are always better than that of single SVM with different combinations of feature subsets. One possible reason is that different feature subsets can provide sufficient diversity among base classifiers of ensemble scheme, individual SVM performances do not relate well to combined performances as they miss out the important information about the team strength of the classifiers. Also can be seen in Figure 2, the accuracy of DT is better than that of Majority Voting scheme with each combination of feature subsets. It may due to that DT and Majority Voting are two different types of ensemble, the former is definite as “classifier fusion” while the latter as “classifier selection” [13], thus in our task DT combines classifier outputs by comparing them to a characteristic template for each accent, and it uses all classifier outputs to calculated the final result, which is sharp contrast to Majority Voting method which only use selected classifier outputs to make final decision.

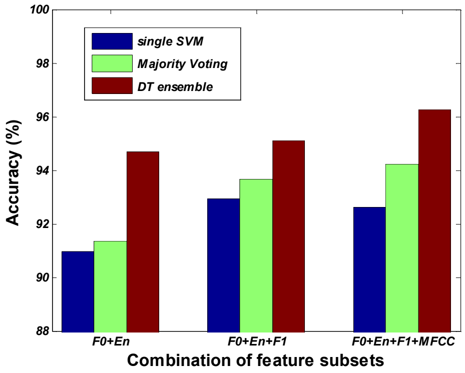

Figure 2: *Comparison of three classification methods: DT ensemble, Majority Voting and single SVM.*

6.3. Results of diversity analysis

As can be seen in Figure 3, the higher diversity is almost reached by DT ensemble scheme with different combinations of feature subsets. This result shows the similar trend with accuracy in the same case. Therefore, there might be the main reason why DT ensemble can produce higher accuracy than Majority Voting ensemble, in term of diversity analysis. However, one of our assumptions is that there also might be some connection between the accuracy and diversity for a certain ensemble scheme with different combinations of feature subsets. Unfortunately, the result shows that there is no obvious relation supporting this assumption. One possible reason is that our ensemble schemes are more sensitive to variations in the number of base classifier than the diversity among different feature subsets.

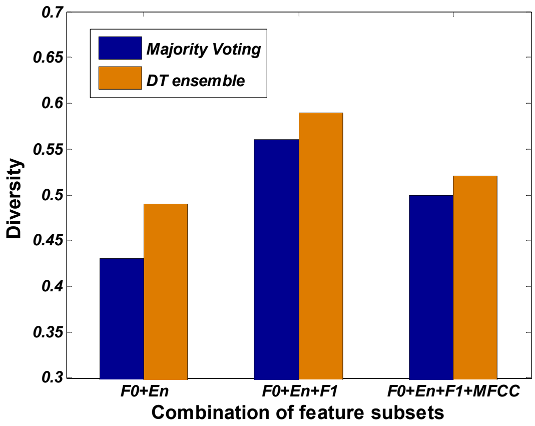

Figure 3: *Comparison of diversity by the entropy measure: DT ensemble, Majority Voting.*

7. Conclusions

This paper proposes a novel scheme for accent classification. A multi-accent Mandarin corpus was adopted for our study, including a standard Mandarin corpus and a minority accent corpus which collected by our laboratory, and we extracted three typical minorities' accents from it. We use decision templates algorithm to ensemble base classifiers built on different acoustic feature subsets. Experiments show that the highest performance of classification is achieved by our ensemble scheme with comparison of Majority Voting and single-based SVM (also as the base classifiers of the ensemble systems). The results also show that the combination with more feature subsets does not always outperform the less one. On the other hand, diversity analysis indicates that the higher diversity is almost reached by DT ensemble scheme with different combinations of feature subsets. This result shows the similar trend with accuracy in same case. But the diversity-analysis results also show there is no obvious relation between the accuracy and diversity for a certain ensemble scheme with different combinations of feature subsets.

However, our experiments merely consider the acoustic-level features of accent issue, linguistics and lexical-level features need to be adopted in future work.

8. Acknowledgements

We gratefully acknowledge the supports from grants of The Applied Basic Research Programs of Yunnan Province (No. 2007F169M) and of The Midlife and Young Backbone Teacher's Training Programs of Yunnan University (2006.1-2008.12). The authors also wish to acknowledge Eric Chang for providing the set of CD-ROMs of speech corpora. Correspondence should be addressed to Jian Yang.

9. References

[1] Arslan, L. M. and Hansen, J. H. L., “Language accent classification in American English ", Speech Communication, 188:353–367, 1996.

[2] Chen, T., Huang, C., Chang, E. and Wang J., “Automatic accent identification using Gaussian mixture models”, IEEE workshop on ASRU., 343- 346, 2001.

[3] Tang, H. and Ghorbani, A. A., “Accent classification using Support Vector Machine and Hidden Markov Model”, Advances in Artificial Intelligence, vol. 2671, pp. 629-631, 2003.

[4] Lincoln, M., Cox, S. and Ringland, S., “A comparison of two unsupervised approaches to accent identification”, in Proc. of ICSLP., 1998.

[5] Ishi, C. T., Hirose, K., Minematsu, N., “More F0 representation for accent type identification in continuous speech and considerations on its relation with perceived pitch values”, Speech communication, vol. 41 pp. 441-453, 2003.

[6] Suomi, K., Toivanen, J., Ylitalo, R., “Durational and tonal correlates of accent in Finnish”, Journal of Phonetics, 31, pp. 113-118, 2003.

[7] Banfield, R. E. et al, “Ensemble diversity measures and their application to thinning,” Information Fusion, vol. 6, pp. 49-62, 2005.

[8] Shipp, C. A. and Kuncheva, L. I., “Relationships between combination methods and measures of diversity in combing classifiers,” Information Fusion, vol. 3, pp. 135-148, 2002.

[9] Yang, J., Wei, H., Pu, Y. and Zhao, Z., “Comparison of non-native speaker adaptations for large vocabulary continuous mandarin speech recognition”, International Journal of Information Technology, vol. 11, pp. 9-19, 2005.

[10] Kuncheva, L. I., “Decision templates for multiple classifier fusion: an experimental comparison,” Pattern Recognition, vol. 34, pp. 299-314, 2001.

[11] Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E., “Support vector machine for speaker and language recognition”, Computer Speech and Language, vol.20, pp. 210-229. 2006.

[12] Schuller, B. and Rigoll, G., “Timing levels in segment-based speech emotion recognition,” Proc. INTERSPEECH-ICSLP, Pittsburgh, USA, 2006.

[13] Tsymbal, A., Pechenizkiy, M., Cunningham, P., “Diversity in search strategies for ensemble feature selection”, Information Fusion, vol. 6, pp. 83-96, 2005.

[14] Morrison, D., Liyanage C. De Silva, “Voting ensemble for spoken affect classification,” Journal of Network and Computer Application, vol. 30, pp.1356-1365, 2007.

[15] Deshpande, S. Chikkerur, S. Govindaraju, V., “Accent classification in speech”, Proceedings of the Fourth IEEE Workshop on Automatic Identification Advanced Technologies, pp. 139-143, 2005.

[16] Yi, L., Pascale, F., “Partial change accent models for accented Mandarin speech recognition”, IEEE Workshop on Automatic Speech Recognition and Understanding, pp.111-116, 2003.