Decision Templates Ensemble and Diversity Analysis for Segment-Based Speech Emotion Recognition

Fukun Bi Jian Yang Ying Yu Dan Xu

School of Information Science and Engineering, Yunnan University
Kunming, 650091, P. R. China

Abstract

In this paper, we propose a novel scheme for speech emotion recognition, which uses decision-templates ensemble algorithm (DT) to combine base classifiers built on segment-level feature sets. Different feature sets from segments 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 at suitable segment levels. On the other hand, we investigate which segment-level and strategy of training base classifiers can provide potential performance in speech emotion recognition, in terms of diversity analysis.

Keywords: Speech emotion recognition, Decision templates ensemble, Diversity analysis, Information fusion, Ensemble classifiers.

1. Introduction

A growing interest in emotion recognition by speech is one of the active research fields for human-computer interaction. In recent works, it is believed that speech emotion recognition systems can obtain better performance when combined with segment-level information besides utterances-level [1] [2] [3]. Nevertheless, most of them consider segment features as the supplement of utterance-feature vector, as in [1] [2]. In other words, a super-vector is constructed for classification by fusion segments and utterance features. Classifier ensemble is a popular approach to improve the performance of recognition systems. Unlike relative works that ensemble different base classifiers (trained with same utterance features) for emotion recognition systems as in [8] [12], we use an effective approach to utilize segment information, which ensembles base classifiers built on segment-level feature sets. Different feature sets from segments can provide sufficient diversity among base classifiers of ensemble.

For these motivations, a novel ensemble scheme based on segment-level features is proposed for speech emotion recognition. We use decision-templates algorithm (DT) to combine homogeneous base classifiers (support vector machine) in ensemble strategy. This ensemble algorithm yields generally better performance among similar algorithms in other pattern recognition tasks [4]. In view of various representations of speech in our case, we propose four strategies to train base classifiers of DT scheme. We focus on the relative time interval segment approach (RTI) which usually outperforms other automatic-segment approaches [1] and can provide fixed number of segment-feature vectors to establish base classifiers for ensemble. On the other hand, it is well-known that diversity is closely related to ensemble systems [9] [10]. 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 scheme for speech emotion recognition in term of diversity analysis. In this paper, we use diversity analysis with the entropy measure (Ent) to discuss which segment-level of RTI and strategy of training base classifiers (we propose) can provide potential performance in our proposed scheme.

The rest of this paper is organized as follows: In Section 2 we introduce an emotion database adopted in our experiments. Section 3 discusses the extraction and selection 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. Emotional Speech Database

We have used Berlin Emotion Speech Database (EMO-DB) in our experiments [5]. The database contains utterances of both male and female speakers. 5 male and 5 female native actors uttered 10 sentences in German that have little emotional content textually. It contains 7 emotion classes: anger, fear, joy, sadness, disgust, boredom and neutral emotion. The complete database was evaluated in a perception test. All final 535 utterances were recorded under studio conditions with high-quality recording equipments and saved in mono wave files with 16,000 Hz sample rate and 16 quantitative bits.

3. Feature Extraction

Acoustic features are widely observed to carry the most significant characteristics of emotion in speech [6] [13]. In this study, we estimated the following acoustic features: pitch (F0), energy, first formant (F1), melfrequency cepstral coefficients (MFCCs) and rhythm, and calculated their correlative statistical features. However, we carried out the investigation at both segment and utterance levels. So we extracted these acoustic features at both levels.

3.1. Segment and utterance representation of speech

Segment-level information is believed to improve the performance in speech emotion recognition besides utterance-level [1] [2] [3]. Different representation of speech constitutes different dataset for our ensemble scheme and diversity analysis. Automatic, word and syllable segment are main methods in previous studies [1] [2] [3]. We focus on automatic segment without considering extra-effort in word or syllable boundary detection with automatic speech recognition. Relative time intervals (RTI) used in our scheme, usually outperforms other automatic segment methods, is proposed in [1] and is defined as splitting an utterance at fixed relative positions, e.g. halves or thirds.

3.2. Acoustic features

We extracted the following acoustic features for RTI segment level in the same manner as whole 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 the energy 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 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. For each level representation, the first 7 MFCCs were extracted from every frame, and the mean were calculated. In addition, we calculated rhythm features: speech-rate, voiced-to-unvoiced region ratio.

Finally, the acoustic-feature vector for each level of speech consists of 36 features, among which 9 from F0, 9 from F1, 9 from energy, 7 from MFCCs and 2 from rhythm.

3.3. Feature subset selection

To eliminate irrelevant acoustic features in the base feature set and improve the performance of classification, feature set selection technique is applied. We use genetic algorithm (GA) to search the optimum feature subset, which has been shown to be an effective global optimization technique [7]. In order to simplify whole scheme and to emphasize our ensemble and diversity-analysis experiments, we focus on the selection of feature subset of utterance-level in EMO-DB instead of both utterance and segment levels.

4. Classification by decision templates ensemble

In pattern recognition systems, it is believed that ensemble techniques have better potential for improvement on accuracy than single-based classifiers. They are widely used in many pattern recognition tasks as in [11], also are they used successfully in speech emotion recognition in recent works [8] [12]. In this paper, we apply decision templates (DT) method in ensemble strategy, which was proposed by L. I. Kuncheva in [4] and was reported a simple and effective method. Especially, it could yield better performance than other similar ensemble schemes.

4.1. Base classifiers of ensemble

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 emotion recognition as in [1] [14].

4.2. Training strategies for base classifiers

Let $\{D_1, D_2, \ldots, D_L\}$ be a set of base classifiers, the number of them is equal to that of RTI segments of an utterance. For training or test phase of our ensemble scheme, they were trained by different feature vectors. We propose four strategies for training these base classifiers. The first strategy (S-U) is defined as following:

- For the training phase, base classifiers are trained with the corresponding segment-feature vectors (e.g. the vector from the *i*-th segment training D_i).
- For the test phase, base classifiers are trained with same utterance-feature vectors.

Other strategies are S-S, U-S and U-U, which are defined in the similar manner as S-U.

4.3. Estimation of decision templates

Each base classifier (trained) gets as its input a feature vector **x,** which comes from the corresponding RTI segment. For example, the D_i gets its input feature vector from the *i*-th segment. Let $\{w_1, w_2, \ldots, w_7\}$ be the set of class labels represent 7 emotion states , we assume that the RTI segments are denoted by the same label with the whole utterance to which they belong. The classifier outputs can be organized in a decision profile (*DP*) as following matrix.

$$DP(\mathbf{x}) = \begin{bmatrix} d_{1,1}(\mathbf{x}) & \ldots & d_{1,j}(\mathbf{x}) & \ldots & d_{1,7}(\mathbf{x}) \\ \ldots & & & & \\ d_{i,1}(\mathbf{x}) & \ldots & d_{i,j}(\mathbf{x}) & \ldots & d_{i,7}(\mathbf{x}) \\ \ldots & & & & \\ d_{L,1}(\mathbf{x}) & \ldots & d_{L,j}(\mathbf{x}) & \ldots & d_{L,7}(\mathbf{x}) \end{bmatrix}. \qquad (1)$$

Where, $d_{i,j}(\mathbf{x})$ is the degree of "support" given by D_i, for the hypothesis the given input **x** comes from the segment labeled w_j, i=1,2,…,L, j=1,2,…7. "Crisp" class labels are adopted in this work based on the outputs of SVM base classifiers. For example, $d_{i,j}(\mathbf{x})$=1, if D_i recognizes correctly, otherwise, $d_{i,j}(\mathbf{x})$=0.

Let Z = $\{z_1, \ldots, z_j, \ldots, z_n\}$ be the labeled data set (utterance level) for EMO-DB. The *i*-th row of $DP(z_j)$ is evaluated by the *i*-th segment of z_j. The decision template *i* (DT_i) for emotion class *i* is the expectation of the *DP*s which are evaluated by training utterances labeled as class *i*.

$$DT_i = \frac{1}{N_i} \sum_{z_j \in w_j \in Z} DP(z_j) \text{ , } i=1,2,\ldots,7, \qquad (2)$$

where N_i is the number of these training utterances. Thus we obtain 7 *DT*s denote 7 emotion states with all training utterances of the database.

4.4. Classification with decision templates

In the test phase, for test utterance z_j, we use its RTI segment features to calculate the $DP(z_j)$. μ_i (z_j) is defined as the similar degree between the current $DP(z_j)$ and DT_i , which is calculated by the Euclidean distance.

$$\mu_i(z_j) = \sum_{j=1}^{7} \sum_{k=1}^{L} (d_{k,j}(z_j) - dt_i(k,j))^2 \qquad (3)$$

Where, $dt_i(k, j)$ is the *k*, *j*-th entry in DT_i. Thus, a test utterance z_j is assigned class label w_k, when μ_k (z_j) is the smallest value among $\{\mu_1(z_j), \mu_2(z_j), \ldots, \mu_7(z_j)\}$.

5. Diversity Analysis

Diversity among base classifiers is recognized as one of the important characters in classifier ensemble scheme [9] [10]. The general motivation is that diversity analysis is expected to achieve higher ensemble accuracy, as in [10]. We are interested not only in the above motivation, but also in the following two comparative experiments:

- We investigate whether there is any connection between accuracy and diversity at different RTI segment levels. For this hypothesis, we use diversity analysis to compare the diversity among segment-classifiers which are built on different segment-feature sets.
- The aim of another hypothesis is to investigate which strategy of training base classifiers (we proposed in section 4) has better potential for ensemble performance in terms of diversity analysis.

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 [10] and defined as:

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

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. More details about this measure can be obtained in [10].

6. Experimental results

Male and female data are considered separately in this work. In all classification and diversity analysis experiments, we employ the 10-fold cross-validation technique (only at utterance level). The whole database is equally divided into 10 subsets. In each round, we use 9 subsets for training and the remainder one for test, and this process is repeated for ten times.

All final results, by the 10-fold cross-validation, are mean of male and female results.

6.1. Results of DT ensemble classification

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 scheme. The accuracy of this classification method is only calculated at utterance-level for comparison.
- Majority voting is one of the most popular techniques used in classifier fusion. It is also applied to relative works [8] [12]. To compare the performance of DT ensemble, we also use segment-feature sets to train base classifiers (SVM) for majority voting scheme at different RTI segment level.

Fig. 1 shows the accuracy comparison among above three classification schemes, and that of four strategies for training base classifiers in our ensemble scheme at different RTI segment levels.

As can be seen in Fig. 1, the accuracy of DT is better than that of majority voting scheme at each RTI segment level, and better than that of single SVM classifier at first four levels. Note that the ensemble schemes do not always outperform the single-based classifiers, even as their base classifiers, in terms of RTI segment method. The highest performance of classification is achieved by DT ensemble scheme. To compare different strategies of training base classifiers in DT ensemble, we observe that S-U is the best on average. This is likely due to the fact that segment-feature sets can provide sufficient information for training base classifiers in the training phase and that training base classifiers with utterance-feature sets can boost the robustness of recognition in test phase.

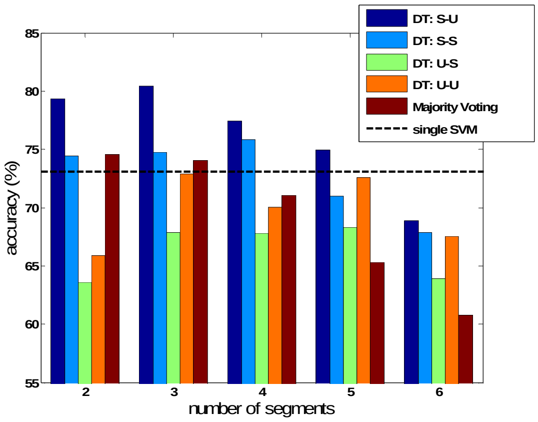

Fig. 1: Comparison of three classification methods: DT ensemble, majority voting and SVM single-based classifier. Comparison of S-U, S-S, U-S, and U-U. Optimal feature vector by GA, EMO-DB.

To better investigate the relationship between the length of RTI segments and the accuracy of DT ensemble scheme, we also compare the accuracy of S-U strategy (always has the best performance in above experiments) at different RTI segment levels. The results are showed in Table 1. The third and the fourth rows in this table respectively represent the average segment-length and standard deviation of whole EMO-DB at different RTI segment levels. For comparison, the results of utterance-level by single SVM classifier are also showed in Table 1 (the second column). As can be seen, the highest accuracy is obtained at the 3-segment by RTI method. We note that not all segment levels of RTI (by DT ensemble) can provide a better performance than utterance level (by single SVM classifier), even worse. Bold numbers in this table represent better results. This means that the segment-based method does not always provide useful information in emotion speech recognition, especially when utterance is split into more segments.

segment level	Utr.	2	3	4	5	6
Accuracy (%)	73.1	**79.3**	**80.5**	**77.5**	**74.9**	68.9
average length(sec)	2.80	1.40	0.93	0.70	0.56	0.47
standard deviation	1.03	0.52	0.34	0.26	0.21	0.17

Table 1: Comparison of Accuracy and segment-length by DT at different RTI segment levels.

6.2. Results of diversity analysis

Fig. 2 shows the diversity comparison at different RTI segment levels, in terms of four strategies for training

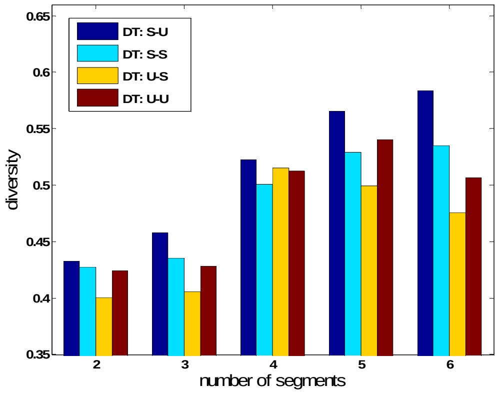

Fig. 2: Comparison of diversity by the entropy measure. Four strategies for training base classifiers of DT ensemble scheme at different RTI levels: S-U, S-S, U-S, U-U. Optimal feature vector by GA, EMO-DB.

base classifiers. As can be seen, the highest diversity is almost reached when the S-U strategy is at the same segment level. This result shows the similar trend with accuracy in the same case. Therefore, there might be a relation between accuracy and diversity at same segment level. However, one of our assumptions is that there also might be some connection between the accuracy and diversity at different RTI segment levels. Unfortunately, the result shows that there is no obvious relation supporting this assumption. One possible reason is that our ensemble scheme is more sensitive to variations in the number of RTI segments than the diversity at different segment levels.

7. Conclusions

This paper proposes a novel scheme for speech emotion recognition. We use decision templates algorithm to ensemble base classifiers (SVM) built on segment-feature sets. Experiments on EMO-DB 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 DT ensemble). The results also show that the RTI segment-based method does not always provide useful information in emotion speech recognition, especially when utterance is split into more segments. On the other hand, diversity analysis indicates that S-U strategy has the highest diversity for training base classifiers of DT, which might be the main reason why S-U can produce highest accuracy than other three strategies at the same RTI segment level. But the diversity-analysis results also show that there is no obvious relation between the accuracy and diversity at different RTI segment levels. One possible reason is that our ensemble scheme is rather sensitive to variations in the number of segments.

In our future work we plan to investigate different segment-based methods by our scheme, in order to test its robustness for speech emotion recognition.

Acknowledgment

This work is partially supported by National Natural Science Foundation of China (No: 60265001). The authors are grateful for using EMO-DB provided by TU Berlin. Correspondence should be addressed to Jian Yang.

References

[1] B. Schuller and G. Rigoll, Timing Levels in Segment-Based Speech Emotion Recognition, *Proc. INTERSPEECH-ICSLP*, Pittsburgh, USA, 2006.

[2] M. Rotaru and D. J. Litman, Using Word-level Pitch Features to Better Predict Student Emotions during Spoken Tutoring Dialogues, *Proc. INTERSPEECH-EUROSPEECH*, Lisbon, Portugal, 2005.

[3] T. Vogt and E. Andre, Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition, *Proc. INTERSPEECH-ICME,* Amsterdam, Holland, 2005.

[4] L. I. Kuncheva, Decision Templates for Multiple Classifier Fusion: An Experimental Comparison, *Pattern Recognition*, vol. 34, pp. 299-314, 2001.

[5] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss, A Database of German Emotional Speech, *Proc. INTERSPEECH-ISCA*, Lisbon, Portugal, 2005, pp. 1517-1520.

[6] O.W. Kwon, K. Chan, J. Hao and T. W. Lee, Emotion Recognition by Speech Signals, *Proc. EUROSPEECH,* Geneva, Switzerland, 2003.

[7] H.Frohlich, O.Chapelle, B. Scholkopf, Feature Selection for Support Vector Machines by Means of Genetic Algorithm, *IEEE Trans. on ICTAI,* Tubingen, Germany, 2003.

[8] D. Morrison, R. Wang, W. L. Xu, Voting Ensemble for Spoken Affect Classification, *Journal of Network and Computer Application,* in press.

[9] R.E. Banfield, et al, Ensemble Diversity Measures and Their Application to Thinning, *Information Fusion*, 6: 49-62, 2005.

[10] C. A. Shipp and L. I. Kuncheva, Relationships between Combination Methods and Measures of Diversity in Combing Classifiers, *Information Fusion,* 3: 135-148, 2002.

[11] J. Czyz, M. Sadeghi, J. Kittle, L. Vandendorpe, Decision Fusion for Face Authentication, *Proc. First International Conference on Biometric Authentication,* Hong Kong, pp. 686-693 , 2004.

[12] D. Morrison, R. Wang, L. C. D. Silva, Ensemble Methods for Spoken Emotion Recognition in Call-Centre, *Speech Communication,* 49: 98-112, 2007.

[13] J. Liscombe, J. Venditti and J. Hirschberg, Classifying Subject Ratings of Emotional Speech Using Acoustic Features, *Proc. INTERSPEECH-EUROSPEECH*, Geneva, Switzerland, 2003.

[14] H. Hao, X. Ming-Xing, W. Wei, GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition, *Proc. IEEE International Conference on ICASSP*, Honolulu, HI, 2007.