Speech Emotion Recognition Based on a Fusion of All-Class and Pairwise-Class Feature Selection

Jia Liu¹, Chun Chen¹, Jiajun Bu¹, Mingyu You¹, and Jianhua Tao²

¹ College of Computer Science, Zhejiang University, Hangzhou, P.R. China, 310027
{liujia, chenc, bjj, roseyoumy}@zju.edu.cn

² National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, P.R. China, 100080
jhtao@nlpr.ia.ac.cn

Abstract. Traditionally in speech emotion recognition, feature selection(FS) is implemented by considering the features from all classes jointly. In this paper, a hybrid system based on all-class FS and pairwise-class FS is proposed to improve speech emotion classification performance. Besides a subset of features obtained from an all-class structure, FS is performed on the available data from each pair of classes. All these subsets are used in their corresponding K-nearest-neighbors(KNN) or Support Vector Machine(SVM) classifiers and the posterior probabilities of the multi-classifiers are fused hierarchically. The experiment results demonstrate that compared with the classical method based on all-class FS and the pairwise method based on pairwise-class FS, the proposed approach achieves 3.2%-8.4% relative improvement on the average F1-measure in speaker-independent emotion recognition.

1 Introduction

Emotion recognition is one of the latest challenges in intelligent human-machine communication. It has broad applications in areas such as customer service, medical analysis and entertainment, etc. As a major indicator of human affective states, speech plays an important role in detecting emotions[1]. Speech emotion recognition can be formulated as three basic steps: extracting acoustic features from speech signals, selecting a feature subset and detecting emotions with various classifiers. Among them, feature selection(FS) is a significant step since it can decrease computational complexity and prediction error by avoiding the 'curse of dimensionality'[2].

In traditional speech emotion recognition, features from all classes are analyzed simultaneously to make a lower-dimensional set which is subsequently used in training/testing an appropriate classifier. Chuang and Wu[3] adopted Principal Component Analysis(PCA) to reduce dimensionality of features for classifying seven basic emotions. Ververidis et al.[4] used Sequential Forward Selection(SFS) to choose five best features from speech signals for recognizing five emotional states. All of them considered the data from all classes together. But intuitively, the features useful to distinguish the emotion 'Angry'

from 'Happy' may differ from those features distinguishing 'Angry' from 'Sad'. In other words, the feature subset produced by all-class FS is not optimal in discriminating between a specific pair of classes. In contrast to the all-class method, a pairwise-class FS has been proposed for musical instrument classification[5] and face recognition[6]. They suggested choosing the most efficient feature subset for each pair of classes. Every subset was dealt with a one-against-one classifier correspondingly and then a 'majority voting' was applied on multi-classifiers to get a final decision. It was reported that compared with the classification based on all-class FS, a relatively high recognition accuracy was achieved by the pairwise method. However, little work of speech emotion recognition has focused on the pairwise framework.

In this paper, we analyze the data by all-class FS and pairwise-class FS respectively. Based on the two views, a hybrid speech emotion recognition system is developed. In the experiments, KNN and SVM are both employed to demonstrate the performance of the novel system.

2 Emotion Corpus and Acoustic Features

The speech corpus contains six emotions: *Neutral*, *Anger*, *Fear*, *Happiness*, *Sadness* and *Surprise*. Three hundred sentences, each expressed in the six categories, are recorded from four Chinese native speakers(two females and two males). In our experiments, speaker-independent emotion recognition of each gender was investigated. On each gender, 80% of utterances were randomly selected for training, and the remaining 20% for testing. The entire experiment was repeated ten times with different training and testing data sets, and the results were averaged. The classification performance was evaluated in terms of the F1-measure, which is defined as $F1 = 2 \times precision \times recall \div (recall + precision)$. The recall is the portion of the correct categories that were assigned, while the precision measures the portion of the assigned categories that were correct.

In the field of speech emotion recognition, many acoustic features have been presented in the literature[4,7]. We extracted 16 formant frequency features and 48 prosodic features from signals since formant frequency features are widely used in speech processing applications and prosody is believed to be the primary indicator of human affective states. The detailed features are indexed as follows:

- 01.-04. *max, min, mean, variance* of the first formant
- 05.-08. *max, min, mean, variance* of the second formant
- 09.-12. *max, min, mean, variance* of the third formant
- 13.-16. *max, min, mean, variance* of the fourth formant
- 17.-20. *max, min, mean, median value* of pitch
- 21.-22. *mean, median value* of pitch rises
- 23.-25. *max, mean, median duration* of pitch rises
- 26.-27. *mean, median value* of pitch falls
- 28.-30. *max, mean, median duration* of pitch falls
- 31.-32. *mean, median value* of plateaux at pitch maxima[4]
- 33.-35. *max, mean, median duration* of plateaux at pitch maxima

- 36.-37. *mean, median value of plateaux at pitch minima*
- 38.-40. *max, mean, median duration of plateaux at pitch minima*
- 41.-44. *max, min, mean, median value of energy*
- 45.-46. *mean, median value of energy rises*
- 47.-49. *max, mean, median duration of energy rises*
- 50.-51. *mean, median value of energy falls*
- 52.-54. *max, mean, median duration of energy falls*
- 55.-56. *mean, median value of plateaux at energy maxima*
- 57.-59. *max, mean, median duration of plateaux at energy maxima*
- 60.-61. *mean, median value of plateaux at energy minima*
- 62.-64. *max, mean, median duration of plateaux at energy minima*

3 Experiment Design

3.1 System Overview

The novel system of speech emotion recognition takes account of the most relevant features from all-class data and the optimal feature subsets from all possible pairs of classes. Mainly, the system is composed of two phases: training(Fig. 1) and testing(Fig. 2). Each phase can be divided into two components: an all-class structure and a pairwise-class structure.

In the off-line training phase, the all-class structure analyzes the data from six classes jointly and produces a selected feature subset. It outputs a classification model based on the subset with its label information and creates a feature index table which is an array of selected feature indices. On the other hand, a parallel pairwise-class structure is implemented. For each pair of classes, feature selection

Fig. 1. The framework of the training phase

Fig. 2. The framework of the testing phase. The rounded rectangles with dot lines are the acquisition of the training phase.

and corresponding model training are performed. Supposing N is the number of distinct classes, the pairwise-class structure makes $N(N-1)/2$ feature index tables and $N(N-1)/2$ classification models for later testing.

While testing with the 64-dimensional data, FS runs $N(N-1)/2 + 1$ times according to the different feature index tables. Afterwards, the subset selected by the all-class feature index table is recognized by the model obtained from the all-class structure of training. In this process, an estimate of posterior probabilities rather than only predicting a class label, is required because of a final combination with the pairwise-class structure. Moreover, in the pairwise-class structure, $N(N-1)/2$ one-against-one classifiers implement recognition on their corresponding feature subsets side by side and output $N(N-1)/2$ sets of posterior probabilities. These sets are fused by a multi-classifiers combining strategy to attain the whole posterior probabilities of the pairwise-class structure. Ultimately, two sets of posterior probabilities from the all-class structure and the pairwise-class structure are combined to determine the classification result.

3.2 Feature Subset Selection

SFS, the first algorithm[8] proposed for feature selection, is simple and empirically successful. It starts with an empty subset of features and performs a hill-climbing deterministic search. In each iteration, the feature not yet selected is individually incorporated in the subset to calculate a criterion. Then the feature which yields the best criterion value is included in the new subset. This iteration will not be stopped until no improvement of the criterion value is achieved.

Fig. 3. Indices of 16 feature subsets selected by SFS with KNN on female utterances. X-axis is the indices of all 64 features(see details in Section 2), and Y-axis represents the involved classes. The corresponding words of abbreviated letters are as follows: All-All six emotions, An-Anger, Fe-Fear, Ha-Happiness, Ne-Neutral, Sa-Sadness, Su-Surprise.

The criterion employed to guide the search was the average classification F1-measure by the selected features. The F1-measure was calculated by a 5-fold cross-validation with certain classifier. Fig.3 shows the array elements of some feature index tables selected by SFS with KNN on female utterances. These subsets were quite distinct since they were selected on different involved classes. The SFS with KNN on male and SFS with SVM have also been performed, yet the results are omitted.

3.3 Classifiers and Probability Estimates

In pattern recognition, KNN is the simplest algorithm only based on memory. Given a query sample x , it finds K closest neighbors of x in training nodes by a distance metric(e.g. Euclidean distance in our study) and predicts the class label using majority voting by the labels of the K nodes. In contrast to other statistical classifiers, KNN needs no model to fit. This property simplifies the structures of the training phase by no model training, thus the training phase with KNN classifier only requires selecting features. As a preprocessing of experiments, the value of K was determined by a 5-fold cross-validation on 64-dimensional training data from all classes. It was conducted on the points $K : [10, 15, \dots, 70, 75]$. At each K , the average F1-measure on validation data was obtained from the iterations of the cross-validation. The experiment results in Fig.4(a) show that $K = 35$ performs best on each gender.

Another powerful tool 'SVM', has been originally proposed to classify samples within two classes. It maps training samples of two classes into a higher-dimensional space through a kernel function and seeks an optimal separating hyperplane in this new space to maximize its distance from the closest training point. While testing, a query point is categorized according to the distance between the point and the hyperplane. We employed binary SVM with a radial

Fig. 4. Average classification performance of KNN/SVM on different values of parameters

basis function(RBF) kernel, and for the all-class structure, we extended binary SVM with the 'one-against-one' rule which has offered empirically good performance in multi-class pattern recognition[9]. The classification error rate of a binary RBF-SVM is largely dependent on a kernel parameter γ and a regularization parameter C . We selected them by another 5-fold cross-validation based on the 64-dimensional training data over all classes. The average F1-measures of the experiments on $(\gamma, C) : [1, 0.1, 0.01] \times [100, 10, 1]$ are shown in the Fig.4(b) and we chose (0.1, 10) as the fix value of (γ, C) in our experiments.

For the sake of probability estimates, probabilistic outputs of KNN and SVM were required rather than label prediction. So a well-known estimate of KNN which treats each neighbor equally was adopted[10]: for a query point x , the posterior probability P_i of class i equals the fraction of the K nearest neighbors that belong to class i , i.e. $P_i = K_i/K$ where K_i denotes the number of the nodes belonging to class i in the K neighbors. For the binary SVM, the Platt's posterior probability estimate[11] by a sigmoid function was calculated as follows: given the training examples labelled by $(1, -1)$, the SVM computes a decision function f such that $sign(f(x))$ is used to predict the label of any query sample x . The posterior probability of x belonging to class '1' is: $P(y = 1|x) \approx P_{A,B}(x) \equiv (1 + \exp(Af(x) + B))^{-1}$, where A and B are estimated by minimizing the negative log-likelihood function using the validation data set.

3.4 Fusion Strategy

Once the 15 binary classifiers in the pairwise-class structure have been implemented, the next important task involves the fusion of these individual results. Many multi-classifier combining strategies have been proposed in recognition applications, such as majority voting, sum-rule, product-rule and belief function, etc. The sum-rule, where the continuous outputs like posterior probabilities are summed up, empirically perform well in multi-classifier problems[12]. We adopted it in the fusion of all binary classifiers and denoted the outputs from sum-rule by $P = (p_1, p_2, \dots, p_6)$. For the later combination with the probabilities from the all-class structure, an output square sum normalization[13] of P was required to solve the problem of output incomparability. The normalization

result was denoted by P' , satisfying $\sum_{i=1}^6 p'_i = 1, 0 \leq p'_i \leq 1$. Supposing the posterior probabilities from the all-class structure are $Q = (q_1, q_2, \dots, q_6)$, where $\sum_{i=1}^6 q_i = 1, 0 \leq q_i \leq 1$, the Q and P' were also combined by sum-rule.

4 Performance Comparison

In order to evaluate the system performance, the classical method based on all-class FS and the pairwise method based on pairwise-class FS were investigated, i.e. the all-class structure and the pairwise-class structure were considered separately to do the speech emotion classification. The average F1-measure of three methods on each emotion is shown below. With KNN on female model(Fig.5(a)), it is observed the F1-measure of the pairwise method is better than those of the all-class method on every emotion. But on male model(Fig.5(b)), these two methods achieve comparable performance. Besides, the comparisons using SVM in Fig.5(c)(d) show that the method based on all-class FS performs better than the pairwise method in a majority of emotions on both female and male models. This is contrary to the previous researches in [5,6] and the results by KNN. Nevertheless, our proposal is more stable and it performs substantially better in almost all cases. Compared with the all-class method, our approach using KNN reaches 7.6% improvement on female and 4.4% on male in the average F1-measure of six emotions. And it increases by 3.2%(female) and 3.6%(male)

(a) Female and KNN

(b) Male and KNN

(c) Female and SVM

(d) Male and SVM

Fig. 5. Performance comparison of All-Class(the method based on all-class FS), Pairwise-Class(the method based on pairwise-class FS) and Our Approach

with SVM. The relative improvements to the pairwise method are 3.9%(female) and 4.7%(male) using KNN, 8.4%(female) and 7.7%(male) with SVM.

5 Conclusions and Future Work

This paper proposes a novel speech emotion recognition system based on a fusion of all-class FS and pairwise-class FS. It is a combination of the classical approach which only produces a feature subset in view of all-class data and the pairwise method selecting features on each pair of classes. Experiment results show that our proposal makes relative improvement in almost all cases, compared with the classical method and the pairwise method by the KNN/SVM classifier. In this paper, we have only investigated SFS, KNN/SVM classifiers and audio features. More effort will be paid on the generalization of the hybrid system and multi-modal emotion recognition will be included in the future work.

References

1. A.Mehrabian: Communication without words. *Psychology Today*, **2(4)** (1968) 53-56
2. R.Bellman: *Adaptive Control Processes: A Guided Tour*. Princeton University Press, (1961)
3. Z.J.Chuang, C.H.Wu: Emotion Recognition Using Acoustic Features and Textual Content. *Proc. International Conference on Multimedia and Expo*, **1** (2004) 53-56
4. D.Ververidis, C.Kotropoulos, I.Pitas: Automatic Emotional Speech Classification. *Proc. International Conference on Acoustics, Speech, and Signal Processing*, **1** (2004) 593-596
5. S.Essid, G.Richard, B.David: Musical Instrument Recognition Based on Class Pairwise Feature Selection. *Proc. International Conference on Music Information Retrieval*, (2003) 560-567
6. G.Guo, H.Zhang, S.Li: Pairwise Face Recognition. *Proc. International Conference on Computer Vision*, **2** (2001) 282-287
7. B.Schuller, G.Rigoll, M.Lang: Hidden Markov Model-based Speech Emotion Recognition. *Proc. International Conference on Acoustics, Speech, and Signal Processing*, **2** (2003) 1-4
8. I.Guyon, A.Elisseff: An Introduction to Variables and Feature Selection. *Machine Learning Research*, (2003) 1157-1182
9. N.Cristianini, J.Shawe-Taylor: *An Introduction to Support Vector Machines*. Cambridge University Press, (2000)
10. K.Fukunaga, L.Hostetler: K-nearest-neighbor Bayes-risk Estimation. *IEEE Transactions on Information Theory*, **21(3)** (1975) 285-293
11. J.Platt: *Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods*. MIT Press, (2000)
12. J.Kittler, M.Hatef, R.Duin, J.Matas: On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20(3)** (1998) 226-239
13. Y.S.Huang, C.Y.Suen: A Method of Combining Multiple Classifiers- A Neural Network Approach. *Proc. the 12th IAPR International Conference on Pattern Recognition*, **2** (1994) 473-475