A Hybrid PNN-GMM Classification Scheme for Speech Emotion Recognition

Wee Ser[1], Ling Cen[2], Zhu Liang Yu[1]

[1]Centre for Signal Processing, Nanyang
Technological University, Singapore
{ewser, ezlyu}@ntu.edu.sg

[2] Institute for Infocomm Research
Singapore
lcen@i2r.a-star.edu.sg

Abstract

With the increasing demand for spoken language interfaces in human-computer interactions, automatic recognition of emotional states from human speeches has become of increasing importance. In this paper, we propose a novel hybrid scheme that combines the Probabilistic Neural Network (PNN) and the Gaussian Mixture Model (GMM) for identifying emotions from speech signals. In order to handle mismatches more effectively, the Universal Background Model (UBM) is incorporated into the GMM, and the resultant model is denoted as UBM-GMM. In the hybrid scheme, the strengths of the PNN and the UBM-GMM are combined through a novel conditional-probability based fusion algorithm. Experimental results show that the proposed scheme is able to achieve higher recognition accuracy than that obtained by using PNN or UBM-GMM alone.

1. Introduction

The human-computer interaction technology has advanced rapidly over the recent decade. In particular, the development of conversational interfaces has been playing a critical role in realizing natural and effective communications between human and computers. One important component of conversational interfaces is the recognition and synthesis of human speech signals. It is well known that human speech conveys not only the linguistic content but also the emotion of the speaker. Although the emotion does not alter the linguistic content, it carries important information on the speaker and his/her responses to the outside world [1] [2]. As such, it is important that the computer understands the emotional states conveyed in human speech in human-computer interaction applications.

Automatic identification of emotional states from speech is a challenging task. This is due partly to the lack of a precise definition of emotions [3]. In [3], the acoustic information has been combined with the lexical and discourse information for speech emotion recognition. In that paper, the Linear Discriminant Classifiers (LDC) with Gaussian class-conditional probability and k-nearest Neighborhood classifiers (K-NN) are used for recognizing negative and non-negative emotions from speech signals. In [4], a rough set theory is used for feature selection and the Support Vector Machine (SVM) is employed for emotion classification. Earlier research has also focused mainly on feature extraction, selection and the use of a single classifier for emotion recognition [5]. Although several papers have been published on speech emotion recognition, very few have considered hybrid classification methods [6]. In [6], two hybrid schemes, stacked generalization [7] and the unweighted vote, are applied for emotion recognition. These two schemes have accuracies of 72.18% and 70.54% respectively, when they were used to recognize 6 emotions. In [5], a hybrid classification method that combines the SVM and the Decision Tree (DT) is proposed for emotion recognition. The average accuracies for classifying 4 emotions are reported to be 72.4%, 70.8%, and 71.3%, respectively.

In this paper, we propose a novel hybrid scheme that combines the Probabilistic Neural Network (PNN) and the Gaussian Mixture Model (GMM), for speech emotion recognition. The GMM used is incorporated with the Universal Background Model (UBM) and the resultant model is denoted as the UBM-GMM. An important novelty introduced by the hybrid scheme here is the proposal of a conditional-probability based fusion algorithm, which combines the strengths of the PNN and the UBM-GMM methods effectively.

The remaining part of this paper is organized as follows. The acoustic feature extraction process is discussed in Section 2. In Section 3, the proposed hybrid classification scheme is presented. Numerical

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

results are shown in Section 4 and the concluding remarks are given in Section 5.

2. Acoustic Feature Extraction for Emotion Recognition

Feature extraction is an important process required in the automatic recognition of emotional states from human speeches. The features and process involved are briefly described below.

The speech data are first high-pass filtered by a FIR filter given by

$$H(z) = 1 - 0.9375z^{-1}. \qquad (1)$$

Signal frames of length 25 msec are then extracted from the filtered speech signal at an interval of 10 msec. A Hamming window is applied to each signal frame to reduce signal discontinuity.

Acoustic features include pitch, intensity, and the Mel-Frequency Cepstrum Coefficients (MFCC). It has been shown in the literature that basic acoustic features extracted directly from the original speech signals (e.g. pitch, energy related features) are reliable for emotion recognition [1] [2]. MFCC [8] are features derived from the basic acoustic features and they are used widely in speech recognition. Numerical studies have shown that MFCC can also achieve excellent classification accuracy in speech emotion recognition. The list below shows the feature set used for speech emotion recognition.

1) Pitch: median, mean, standard deviation, maximum, minimum, range (max-min)
2) Intensity: median, mean, standard deviation, maximum, minimum, range (max-min), first quartile, and third quartile
3) Low-passed Intensity (cut-off frequency at 250 Hz): median, mean, standard deviation, maximum, minimum, range (max-min)), first quartile, and third quartile
4) Mel-Frequency Cepstrum Coefficients (MFCC): 24 MFCC coefficients

3. Proposed Hybrid Classification Scheme

Most emotion recognition schemes use a single classifier [6]. In order to improve the overall recognition accuracy, a hybrid scheme that combines the strengths of two classifiers is proposed in this paper.

The structure of the proposed scheme is shown in Fig. 1. In Figure 1, the PNN and UBM-GMM are used as the two base classifiers. The outputs of the two base classifiers are processed by a proposed conditional probability fusion algorithm, before the final decision is made.

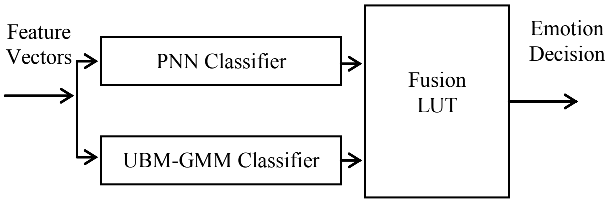

Figure 1. Structure of Proposed Hybrid Scheme (LUT: Look Up Table)

PNN [9] has been employed as a pattern classifier in many applications. The strengths of PNN include simple training process, fast convergent rate, and ease of implementation. It is a Bayesian statistical classifier that uses Parzen estimator to approximate class dependent PDF.

The GMM is basically a single-state HMM (Hidden Markov Model) with a Gaussian mixture observation density. It assumes that the observed variables are generated via a PDF that is obtained by linearly combining a set of weighted Gaussian PDF's. This has been shown to be an effective PDF for text-independent speaker recognition when no *prior* knowledge is available on what the speaker will say [10]. In order to handle mismatches more effectively, the UBM (Universal Background Model) is incorporated into the GMM, and the resultant model is denoted as UBM-GMM. Such a model has been used successfully in speaker verification [10]. In this model, besides using the emotional model λ_{c_i} for the emotional state c_i, a background model $\lambda_{\bar{c}_i}$ is also used to generate the PDF of the feature vectors that do not belong to c_i. Thus, each emotional state is trained with two models, λ_{c_i} and $\lambda_{\bar{c}_i}$ (see Fig. 2). The combined PDF is given as

$$S_{c_i} = \log p\left(X \mid \lambda_{c_i}\right) - \log p\left(X \mid \lambda_{\bar{c}_i}\right). \qquad (2)$$

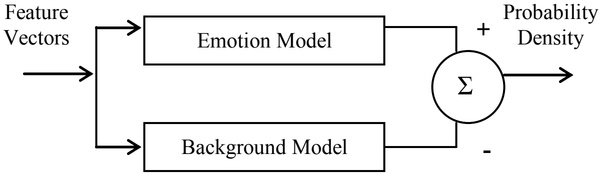

Figure 2. Output Probability Density of UBM-GMM

Training

The proposed hybrid scheme consists of two steps in the training stage. Each step uses a different part of the speech data in the training set. The first step is to train the PNN and the UBM-GMM classifiers individually. In the second step, the confusion matrices for the two base classifiers (i.e. PNN and UBM-GMM) are calculated. A Fusion Look-Up Table (LUT), denoted as, *F*, will also be formed.

Let *L* be the total number of emotional states to be classified. The *L*-by-*L* confusion matrix for each of the base classifier takes the following form:

$$\begin{pmatrix} p_{11} & \cdots & p_{1L} \\ \vdots & \ddots & \vdots \\ p_{L1} & \cdots & p_{LL} \end{pmatrix}$$

In the matrix, $p_{i,j}$ is the probability of the emotional state being c_i given that the estimated emotional state of the classifier being c_j. For an effective classifier, the values of the diagonal entries are expected to be much higher than those of the non-diagonal entries.

The proposed Fusion LUT, *F*, records all possible emotional states estimated by the two classifiers, the actual emotional states, and the conditional probability of the actual emotion being one of the emotional states. This is elaborated below.

Since there are two base classifiers, the number of possible combinations of the states estimated by the 2 classifiers is L^2 . Let c_{PNN} and c_{GMM} be the emotional states estimated by PNN and UBM-GMM respectively, and c_r be the actual emotional state. A typical row of *F* takes the form,

$$[c_{PNN} \quad c_{GMM} \quad c_r \quad p(c_r)\,]$$

where $p(c_r)$ is the conditional probability of the emotional state c_r given by

$$p(c_r) = prob(c = c_r \mid c_{PNN}, c_{GMM}). \qquad (3)$$

In this paper, $p(c_r)$ is approximated as

$$p(c_r) \approx N_{c_{PNN}-c_{GMM}} / N_{c_r}, \qquad (4)$$

where $N_{c_r-c_{PNN}-c_{GMM}}$ represents the number of utterances whose emotional state being c_r given that the estimated emotional state from the PNN and GMM classifiers being c_{PNN} and c_{GMM}, respectively, and N_{c_r} denotes the number of utterances expressed in the emotional state c_r. Note that $N_{c_r-c_{PNN}-c_{GMM}}$ and N_{c_r} count only the utterances used in the second step of the training stage, i.e. calculating the LUT. The dimension of *F* is therefore $L^3 \times 4$. In the ideal situation when the recognition accuracy of every single classifier is 100%, $p(c_r)$ becomes

$$p(c_r) = \left\{ \begin{array}{ll} 1, & \text{for } c_{PNN} = c_{GMM} = c_r \\ 0, & \text{otherwise} \end{array} \right\}. \qquad (5)$$

The advantage of this method is its simplicity, while its disadvantage is the need of a large training dataset.

Testing

During the testing stage, the emotion of a speech sample is determined by either the Fusion LUT or the confusion matrices. Unlike other linear combination methods where the weights need to be adjusted, the proposed fusion algorithm does not need to adjust nor specify any parameter values.

The training-testing process can be summarized into the following 6 steps.

Step 1 Train each of the two classifiers, PNN and UBM-GMM, independently using the training data.

Step 2 Use the trained classifiers to recognize the emotions of another speech data training set**.**

Step 3 Calculate the confusion matrices for both the base classifiers.

Step 4 Calculate the fusion LUT, *F*, according to the process described before.

Step 5 Apply the two base classifiers to the test data separately, and obtain the estimated emotional states, c_{PNN} and c_{GMM}, respectively.

Step 6 Compare the values of $p(c_r)$ in the fusion LUT where the first 2 indices are c_{PNN} and c_{GMM}. Determine the fusion output as $c_{fus} = c_r$ where the value $p(c_r)$ is the highest. In the case when $p(c_r) = 0$ (which can happen when the training sample size is too small), compare the values of $p_{i,i}$ in the two confusion matrices, where *i* corresponds to c_{PNN} and c_{GMM}, for the respective confusion matrices. The final decision of the emotional state, c_{fus} is then taken to be the c_r corresponding to the highest $p_{i,i}$.

4. Experiment

The speech emotion database used in this study is extracted from the Linguistic Data Consortium (LDC) Emotional Prosody Speech corpus (catalog number LDC2002S28), which was recorded by the Department of Neurology, University of Pennsylvania Medical School [11]. It comprises expressions spoken by 3 male and 4 female actors. The speech contents are neutral phrases like dates and numbers, e.g. “September fourth” or “eight hundred one”, which are expressed in the various emotional states as well as neutral state. The number of utterances used is approximately 1200.

Eight emotions, which include anxiety, contempt, despair, disgust, angry, panic, sadness, and neutral, are tested in the experiment. The system is trained in speaker-independent mode. The number of the Gaussian components is chosen to be 64. Half of the data are employed to train the PNN and UBM-GMM classifiers, a quarter of the data are used to calculate the Fusion LUT and the confusion matrices, and the rest are used for testing purpose. Figure 3 shows the results obtained by the three methods (i.e. the proposed scheme, the PNN method alone, and the UBM-GMM method alone). The numerical results are shown in Table 1.

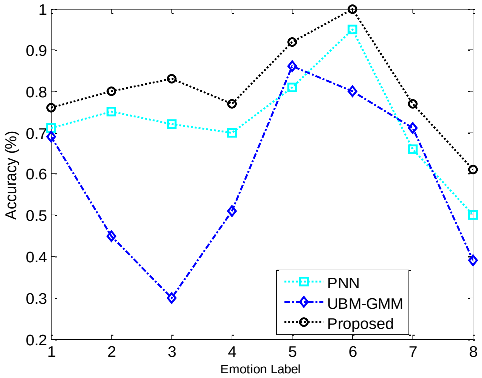

Figure 3. Recognition accuracies of the proposed scheme, PNN and UBM-GMM

Table 1. Recognition accuracies (%) of the proposed scheme, PNN and UBM-GMM

	Proposed	PNN	UBM
Anxiety	76	71	69
Contempt	80	75	45
Despair	83	72	30
Disgust	77	70	51
Angry	92	81	86
Panic	77	66	71
Sadness	61	50	39
Neutral	100	95	80
Average	**80.75**	**72.50**	**58.88**

Figure 3 shows that the proposed hybrid scheme outperforms the PNN and the UBM-GMM methods for all the eight emotional states tested. It can be seen from Table 1 that the proposed scheme is able to achieve 100% accuracy for the neutral state and up to 92% accuracy for the remaining 7 emotional states tested. Compared with the average accuracies of 72.50% and 58.88% achieved by PNN and UBM-GMM alone, the proposed scheme is able to achieve a much higher average accuracy of 80.75%. The confusion matrix and other details / results are included in a journal paper under preparation.

5. Conclusions

This paper presents a novel speech emotion recognition scheme. The proposed scheme introduces a Fusion LUT that effectively combines the strengths of the PNN and the UBM-GMM methods. Experimental results show that the proposed scheme is able to achieve a much higher recognition accuracy compared to that obtained by using PNN or UBM-GMM alone. For the samples tested, the improvement in recognition accuracy is more than 8% when compared to PNN and more than 20% when compared to UBM-GMM.

References

[1] D. Ververidis, and C. Kotropoulos, "Emotional speech recognition: resources, features, and methods," *Speech Communication*, vol. 48, no.9, pp.1163-1181, Sep. 2006.

[2] R. Cowie et al, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32-80, Jan. 2001.

[3] Lee, C., and Narayanan, S., "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp293-303, March 2005.

[4] J. Zhou, G.Y. Wang, Y. Yang, and P.J. Chen, "Speech Emotion Recognition Based on Rough Set and SVM," *Proceedings of 5th IEEE International Conference on Cognitive Informatics*, 2006.

[5] T. Nguyen, I. Bass, "Investigation of Combining SVM and Decision Tree for Emotion Classification," *Proceedings of 7th IEEE International Symposium on Multimedia*, 2005.

[6] D. Morrison, R. Wang, Liyanage C. De Silva, "Ensemble Methods for Spoken Emotion Recognition in Call-centres," *Speech Communication*, vol. 49, pp. 98-112, 2007.

[7] D. H. Wolpert, "Stacked Generalization," *Neural Networks*, vol. 5, pp. 241-260, 1992.

[8] S. B. Davis, P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no.4, pp.357–366, Aug. 1980.

[9] D. F. Specht, Probabilistic neural networks for classification, mapping or associative memory," *Proceedings of IEEE International Conference on Neural Network*, vol. 1, pp. 525-532, 1988.

[10] D. A. Reynolds, "Speaker verification using adapted Gaussian mixture model," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[11] Emotional Prosody Speech Corpus, Linguistic Data Consortium, University of Pennsylvania, PA, USA.