Speech Emotion Recognition Based on SVM and GMM-HMM Hybrid System

Kaiyu Shi[1], Xuan Liu[1], Yanmin Qian[1]

[1]Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
SpeechLab, Department of Computer Science and Engineering
Brain Science and Technology Research Center
Shanghai Jiao Tong University, Shanghai, China

{skyisno.1, liuxuan0526}@gmail.com, yanminqian@sjtu.edu.cn

***Abstract*—Speech emotion recognition is one of the latest challenges in speech processing, we implemented a hybrid system for speech emotion recognition. Two methods were proposed, compared and combined. At first, we utilized GMM-HMM model fed with MFCC feature to exploit the dynamics of emotional signals. Then a balanced SVM classifier was applied on the static LLD feature. In order to take advantage of generative model and discriminate model, these two methods were combined by taking the probability representation of the GMM-HMM model as the additional features for the SVM classifier. This hybrid system makes full use of both MFCC feature and LLD feature, the performance of the hybrid system also proves the effectiveness of our proposed system.**

Key words: speech emotion recognition, SVM, GMM-HMM, hybrid system

1. Introduction

Speech is one of the most effective and convenient approaches for communication. In the last several decades, speech recognition technique is developing rapidly. Aside from vocabulary information in speech, rich information such as emotions, speakers, accents is gradually being utilized.

Speech emotion recognition is one of the latest challenges in speech processing. Besides human facial expressions, speech has proven as one of the most promising modalities for the automatic recognition of human emotions. In the field of security systems, a growing interest can be observed throughout the last year. Besides, the detection of lies, video games and psychiatric aid are often claimed as further scenarios for emotion recognition.[1]

Several previous speech emotion classification studies have proven that speech features and classification models are two most important factors for speech emotion recognition accuracy[2]. The most widely used features are traditional speech recognition features such as MFCC, fbank, and LLD (Low level descriptors). LLD feature is basically a fusion of many ordinary features, such as MFCC[3], zero crossing rate, F0-frequency(pitch), Harmonics to Noise, Root mean square and utterance level feature, in which root mean square is calculated at last over all samples by computing 12 different reduction methods (moments, extremes, linear regression). The features that [4] provides are listed below in table 1. There are various types of classifiers used in the task of speech emotion recognition. Traditional machine learning methods such as HMM, GMM, SVM and k-NN are widely used for long time. Neural Network (NN) is also applied recently to gain good performance.[2]

TABLE 1. Features for the Classifier Sub-Challenge: low-level descriptors (LLD) and functionals.

LLD (16×2)	Functionals (12)
(Δ) ZCR	mean
(Δ) RMS Energy	standard deviation
(Δ) F0	kurtosis, skewness
(Δ) HNR	extremes: value, rel. position, range
(Δ) MFCC 1-12	linear regression: offset, slope, MSE

2. Model Description

2.1 GMM-HMM system

A hidden Markov model is a statistical Markov model in which the system being modeled is assumed to be a Markov chain with unobserved states. A hidden Markov model can be considered as a generalization of a mixture model where the hidden variables, which control the mixture component to be selected for each observation, are related through a Markov process[5], [6].

HMMs are used in speech emotion recognition because speech signals can be described by the transformation of speech states. In a short time-scale, speech can be approximated as a stationary process and be modeled

by Gaussian Mixture Model. Speech can be thought of as a Markov model for many stochastic purposes.

HMMs can be trained automatically and are simple and computationally feasible to use. The hidden Markov model would output a sequence of n-dimensional real-valued vectors (with n being a small integer, for this task, we use MFCC feature, so n is 39), outputting one of these every 10 milliseconds. The vectors consist of cepstral coefficients, which are obtained by taking a Fourier transform of a short time window of speech and decorrelating the spectrum using a cosine transform. The hidden Markov model will tend to have in each state a statistical distribution that is a mixture of diagonal covariance Gaussians. Such distribution assigns likelihood to each observed vector. Then at the ending state, each emotion will have its own output distribution and corresponding likelihood. In this task, a GMM-HMM model consists of 5 states was built for each emotion. The emission density was described by a 30 components Gaussian Mixture Model.

During testing, data point x in the test set is assigned to the label with maximum likelihood:

$$class = argmax_c \log P(x|c) \tag{1}$$

2.2 LLD feature pre-processing

LLD feature is a combination of multiple features, MFCC, zero crossing rate, pitch, Harmonics to Noise, Root mean square and utterance level feature. These features embody useful information related to speech emotion. Different kind of features have different scalars, those features which have large scalar will have larger influence in the classification model. So it is important to normalize the features before training. We standardized the dataset along any axises, centered to the mean and component-wisely scaled to unit variance.

2.3 SVM system

Support vector machines[7] are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each tagged with positive or negative, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

Conventional SVM can only deal with linearly separable problem, kernel trick enables SVM for linearly non-separable problem. Instead of building the nonlinear classification rule of original data points x, SVM with kernel trick learns the corresponding linear rule for the transformed data points $\phi(x)$[8]. Moreover, we are given a kernel function k which satisfies $k(x_i, x_j) = \phi(x_i)\dot{\phi}(x_j)$, the most commonly used kernel is Gaussian radial basis function [9] :

$$k(x_i, x_j) = exp^{(-\gamma\|x_i - x_j\|^2)} \tag{2}$$

RBF kernel is the default kernel, this kernel usually achieves better classification performance than polynomial kernel and linear kernel. For simplicity, we will not emphasize this setting any more.

In this task, we need multi-class classification algorithm. The basic SVM algorithm is a binary classifier. There are two different approaches for SVM to deal with multi-class classification problem. (1) Building binary classifiers which distinguish between one of the labels and the rest (one-versus-all) or (2) Building binary classifiers which distinguish between every pair of classes (one-versus-one). Classification of new instances for the one-versus-all case is done by a winner-takes-all strategy, in which the classifier with the highest output function assigns the class. Since one-versus-one strategy requires more classifiers than one-versus-all version, we used one-versus-all strategy in consideration of speed and overfitting.

The metrics are precision, recall, fscore averaged over classes. In order to obtain high average-precision and average-recall, it becomes necessary to balance the weight of each data point. Otherwise, the classification result will bias towards largest population emotion "neutral", leading to a low average-recall value. The weight of one data point was set inversely proportional to the number of data points in the same class.

2.4 SVM GMM-HMM hybrid system

SVM system and GMM-HMM system use different features and different models, Combining this two models may take advantage of the merit of the two models and yield better classification performance.

For each class c and each data point x, we can calculate the log-likelihood $\log p(x|c)$ from GMM-HMM model. however, we cannot directly use this log-likelihood as feature, because these values is unnormalized and relevant to the time span of x. A typical normalize method is to transform the likelihood towards posterior according to Bayesian equation

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)} \propto p(x|c)p(c) \tag{3}$$

The value $p(c|x)$ is between 0 and 1.

We take these five values (one for each class, sum to 1) as additional features beside LLD, and then train an SVM to classify them.

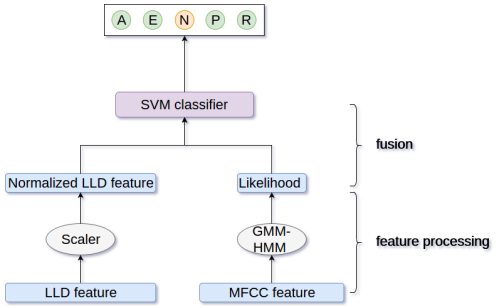

Fig. 1. The architecture of proposed hybrid system

3. Experiments and Analysis

3.1 Dataset and Baseline

We evaluate our model on the 5-class emotion recognition problem in INTERSPEECH 2009 Emotion Challenge (IS2009)[4]. There are 2 problems in IS2009 dataset, 5-class and 2-class emotion classification. For five-class classification problem, "the cover classes **A**nger (subsuming angry, touchy, and reprimanding), **E**mphatic, **N**eutral, **P**ositive (subsuming motherese and joyful), and **R**est are to be discriminated" [4].

We choose the best results reported in [4] as our baseline (precision: 0.3, recall: 0.382; see in Figure 5). The baseline was obtained by an SVM classifier that takes as input the LLD feature preprocessed by SMOTE (Synthetic Minority Oversampling TEchnique) and standardization.

3.2 Data normalization

At the first glance, we plotted a heat map (figure 2) in hopes of getting some basic insights of the LLD feature in the dataset.

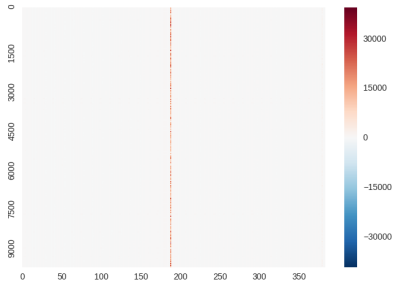

Fig. 2. the heat map of feature value (x-axis: dimensions, y-axis: samples). The background is purely blue except for the vertical line at $x = 189$, it's because the absolute value of 189-th feature is so large that shadows the variances of other features so their colors look like fixed in the figure.

The absolute value of this feature has shadowed the variances of other features so their colors look like fixed in the figure.

After that we decided to take a further step, we plotted the variance and mean of every dimension of features to find out the exact scale relationship among the features.

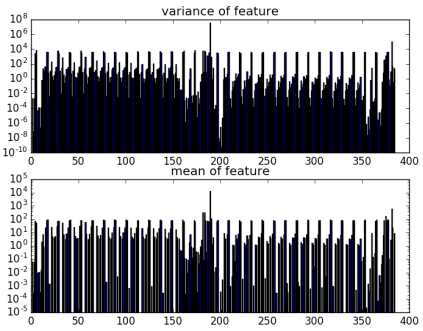

Fig. 3. the logarithm of variance (upper) and mean (lower) with respect to feature dimension. It should be mentioned that the y-axis is in logarithm, so the actual difference of feature scale is much larger than the height difference shown above.

From the figure 2 and figure 3, we can infer that the 189-th dimension of feature is of large scale and changing sharply, the variance of this dimension is 100x as other dimensions of feature. Such kind of huge gap in the range of feature dimensions makes a few dimensions of feature dominate the final classification results, thus exacerbates our classifying task.

To overcome this problem, we exploited feature normalization which is broadly used in data preprocessing step of machine learning. There are many normalization algorithms, sharing the same basic ideas that every dimension of features should be equally important initially. We chose to normalize every dimension to standard normal distribution($\mu = 0, \sigma = I$).

The result is encouraging. In SVM classifier, we obtained a fscore of 0.350 after normalization (Table 2) while the original data can only give us 0.158 (Table 3). The results of original data are extremely rigid since it always predicts 'N' regardless of the inputs, indicating that the classifier is dominated by the class 'N' and its corresponding dimensions of features. Class dominating makes the classifier useless in real world.

TABLE 2. SVM Precision-Recall-Fscore-Support (balanced, normalized)

class	precision	recall	fscore	support
A	0.227	0.463	0.305	611
E	0.393	0.468	0.427	1508
N	0.771	0.560	0.650	5377
P	0.155	0.381	0.220	215
R	0.127	0.181	0.149	546
average	0.335	0.411	0.350	None

TABLE 3. SVM PRECISION-RECALL-FSCORE-SUPPORT (BALANCED, NOT NORMALIZED)

class	precision	recall	fscore	support
A	0.00	0.00	0.00	611
E	0.00	0.00	0.00	1508
N	0.651	1	0.789	5377
P	0.00	0.00	0.00	215
R	0.00	0.00	0.00	546
average	0.130	0.200	0.158	None

TABLE 4. SVM PRECISION-RECALL-FSCORE-SUPPORT (UNBALANCED, NORMALIZED)

class	precision	recall	fscore	support
A	0.555	0.208	0.302	611
E	0.445	0.237	0.309	1508
N	0.695	0.929	0.795	5377
P	0.286	0.047	0.08	215
R	0.5	0.002	0.003	546
average	0.496	0.284	0.298	None

3.3 unbalanced data classification

Another problem in this dataset is data imbalance. Haibo He explained why data imbalance leads to performance corruption, "The fundamental issue with the imbalanced learning problem is the ability of imbalanced data to significantly compromise the performance of most standard learning algorithms. Most standard algorithms assume or expect balanced class distributions or equal misclassification costs. Therefore, when presented with complex imbalanced data sets, these algorithms fail to properly represent the distributive characteristics of the data and resultantly provide unfavorable accuracies across the classes of the data. When translated to real-world domains, the imbalanced learning problem represents a recurring problem of high importance with wide-ranging implications, warranting increasing exploration." [10]

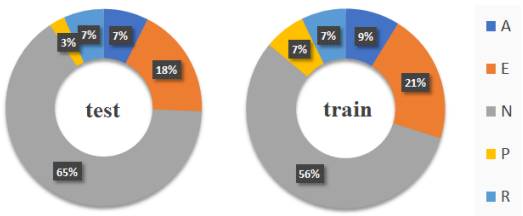

Fig. 4. class compositions of the training set and the test set, class 'N' (gray) takes up the majority of both training and test set

In figure 4 we see that the number of samples in each class is highly imbalanced. Over one-half of samples come from the class 'N', meanwhile the samples of another class are relatively few.

This kind of imbalance among classes may lead to dominant problem too. The more samples one specific class has, the harder the misclassification of this class is to be punished. Then the classifier **prefers class with more samples** [11]. Such property may not cause a problem if we evaluate our final score using a weighted sum among classes. But unfortunately, we have to get the maximum unweighted accuracy in the test set in accordance with [4].

In this paper we used weighted criterion to balance the training samples, which means the weight of each emotion class for SVM training is inversely proportional to the number of samples in each class. It should be noted that we only count the samples in training set. As is shown in figure 5, this trick increased the fscore by about 5% with preprocessed feature in SVM classifier.

$$Loss_{weighted} = \sum_{i \in samples} \frac{count_{all}}{count_i} \times loss_i \qquad (4)$$

3.4 GMM-HMM and SVM Fusion

GMM-HMM is a generative model. For one input sequence of MFCC feature, it calculates the likelihood of each class, then selects the class with maximum likelihood as the final predicted label. GMM-HMM model performs slightly better than baseline (Table 5). We also tried to replace the Maximum Likelihood criterion with Maximum A Posterior criterion, which means, aside from likelihood probability, we also multiply a prior. However, we found no significant differences in overall performance between these two models. We believe that a long time span of the input data makes the prior term much less important than likelihood term.

TABLE 5. GMM-HMM PRECISION-RECALL-FSCORE-SUPPORT

class	precision	recall	fscore	support
A	0.307	0.401	0.348	611
E	0.362	0.528	0.430	1508
N	0.766	0.586	0.664	5377
P	0.147	0.293	0.196	215
R	0.120	0.158	0.136	546
average	0.340	0.393	0.355	None

SVM is a robust discriminant model. After preprocessing the feature and balancing the class weight, SVM achieved good performance as shown in Table 2.

GMM-HMM was trained on continuous MFCC feature, it models the dynamics of emotions. SVM was trained on LLD feature, it models the static overall properties of emotions. These two models are complementary.

Instead of any complicated fusion algorithms, we simply took posterior probability calculated from GMM-HMM model as additional features aside from LLD feature, and fed the extended feature to the support vector

machine (SVM) classifier. This fusion slightly improved the overall fscore by 2 percentage, and the fscores of all the classes except for 'A' were increased (Table 6). We left other fusion methods to our future work.

TABLE 6. SVM AND GMM-HMM HYBRID SYSTEM PRECISION-RECALL-FSCORE-SUPPORT (BALANCED WEIGHT)

class	precision	recall	fscore	support
A	0.270	0.442	0.335	611
E	0.392	0.489	0.435	1508
N	0.772	0.612	0.683	5377
P	0.189	0.353	0.246	215
R	0.149	0.190	0.166	546
average	0.354	0.417	0.373	None

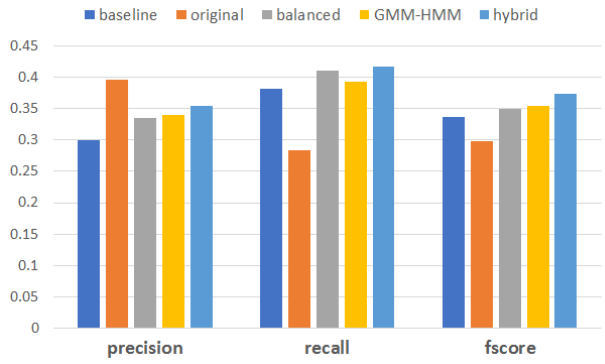

Fig. 5. the performance of SVM with original data, SVM with balanced data, pure GMM-HMM and hybrid system

4. CONCLUSIONS

In this paper, we normalized the LLD feature to eliminate dominate problem, then implemented a hybrid speech emotion detection system. By making full use of the MFCC feature and LLD feature, it combined generative model (GMM-HMM) with discriminate model (SVM). This hybrid system successfully took the advantage of different models and different features, and obtained better performance than both the HMM-GMM model and the SVM model.

5. ACKNOWLEDGEMENT

This work was supported by the Shanghai Sailing Program No. 16YF1405300, the China NSFC projects (No. 61573241 and No. 61603252) and the Interdisciplinary Program (14JCZ03) of Shanghai Jiao Tong University in China. Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

REFERENCES

[1] Á. Urbano Romeu, "Emotion recognition based on the speech, using a naive bayes classifier," B.S. thesis, Universitat Politècnica de Catalunya, 2016.

[2] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[3] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of mfcc," *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[4] B. W. Schuller, S. Steidl, A. Batliner *et al.*, "The interspeech 2009 emotion challenge." in *Interspeech*, vol. 2009, 2009, pp. 312–315.

[5] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[6] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 2. IEEE, 2003, pp. II–1.

[7] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[8] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2001, vol. 1.

[9] A. M. Andrew, "An introduction to support vector machines and other kernel-based learning methods by nello christianini and john shawe-taylor, cambridge university press, cambridge, 2000, xiii+ 189 pp., isbn 0-521-78019-5 (hbk,£ 27.50)." 2000.

[10] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.